Take One Gram of Neural Features,
Get Enhanced Group Robustness

Simon Roburin[*,1,3] Charles Corbière[*,2,3] Gilles Puy[3] Nicolas Thome[2]
Matthieu Aubry[1] Renaud Marlet[3]
Patrick Pérez[3]

[1]LIGM, École des Ponts, [2]Conservatoire National des Arts et Métiers, [3]valeo.ai

Abstract

Predictive performance of machine learning models trained with empirical risk minimization (ERM) can degrade considerably under distribution shifts. The presence of spurious correlations in training datasets leads ERM-trained models to display high loss when evaluated on minority groups not presenting such correlations. Extensive attempts have been made to develop methods improving worst-group robustness. However, they require group information for each training input or at least, a validation set with group labels to tune their hyperparameters, which may be expensive to get or unknown a priori. In this paper, we address the challenge of improving group robustness without group annotation during training or validation. *To this end, we propose to partition the training dataset into groups based on Gram matrices of features extracted by an "identification" model and to apply robust optimization based on these pseudo-groups. In the realistic context where no group labels are available, our experiments show that our approach not only improves group robustness over ERM but also outperforms all recent baselines.*

1. Introduction

Imagine crowd-sourcing an image dataset of camels and cows [4]. Due to selection biases, a high majority of cows stand in front of grass environments and camels in the desert. Therefore, a simple way to differentiate cows from camels would be to classify the background. Such a confounding factor is called a *spurious correlation*. Empirical Risk Minimization (ERM), the most standard machine learning formulation, will naturally exploit this undesirable shortcut and hence perform poorly on minority groups that do not display the same spurious correlation [8,11,28], e.g., a cow standing in the desert. This paper addresses the problem of learning a robust classifier, which would not confuse a cow standing in the desert with a camel, despite having no access to any explicit prior environment knowledge.

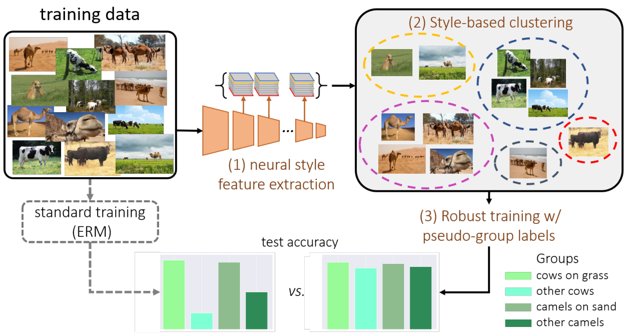

Figure 1. **Overview of the proposed approach for robust classification with unsupervised group discovery.** (1) We first extract deep image features using an identification model and (2) we cluster the training dataset based on their feature Gram matrices (their "style"'); (3) then, we train the targeted classifier with a robust optimization that exploits the assigned pseudo-group labels.

Extensive attempts have been made to develop new training objectives that are robust to spurious correlations, e.g., by ensuring high worst-group accuracy. IRM [3] augments the standard ERM term with invariance penalties across data from different groups. Similarly, [2] promotes, through a simple penalty, identical prediction behaviour across groups. Other works such as [24, 30] minimize explicitly the worst-group loss during training; [25] rebalances majority and minority groups via re-weighting and sub-sampling. However, these approaches require a prior knowledge about the confounding factors during training. This is a major limitation since these factors might be a priori unknown and, if known, ambiguous to define and expensive to annotate.

Recent works [2,6,21,23,27] rely on two-stage schemes, first automatic environment discovery then robust optimiza-

[*] Equal contribution

tion based on environment pseudo-labels. Environment Inference for Invariant Learning (EIIL) [6] derives a group inference objective from a trained *identification model* that maximizes variability across environments, and is differentiable w.r.t. a distribution over group assignments. Just Train Twice (JJT) [21] is a simple method in which environments are defined by images on which a trained identification model performs poorly. GEORGE [27] is based on an unsupervised clustering algorithm in the feature space of a trained identification model. However, all these approaches still require the availability of ground-truth environment labels on a validation set in order to properly tune their hyperparameters.

In the computer vision literature, many identified spurious correlations are closely related to visual aspects, such as background [4], texture [10], image style [13], physical attributes [22] or camera characteristics [16]. In this work, we assume that relevant environment labels can be inferred from visual feature statistics. We propose a two-stage approach, GRAMCLUST, that first assigns a group label, i.e., a class-environment pair label, by partitioning a training dataset into style-based clusters and then trains a robust classifier based on these pseudo-group labels. Our approach is summarized in Fig. 1. The clustering is performed on the Gram matrices of features extracted by an exogenous specifically-trained identification model. Instrumental to the impressive success of style transfer techniques [9], Gram matrices are first and foremost second-order moments of neural activation. Recent work [19] demonstrates that matching Gram matrices is actually equivalent to distribution alignment using the Maximum Mean Discrepancy distance with the second-order polynomial kernel. Therefore, our method can be interpreted as grouping images into clusters of similar feature distributions that are likely candidates for environments. The empirical success of our method on various datasets supports that feature Gram matrices capture more complex visual attributes than just style texture.

Our contributions are: An *easy-to-scale* method to split the training images among distinct pseudo-environments, based on feature Gram matrices; A group-robust learning method, GRAMCLUST, that completely alleviates the need of ground-truth group labels, even in the validation set; Performances on standard image classification datasets with spurious correlations that surpass all recent baselines addressing robustness without group annotation.

2. GRAMCLUST

Our method, GRAMCLUST, is made of two main steps. First, we discover pseudo-environments among the images of a given dataset. Second, we train a robust classifier that leverages the inferred pseudo-environments labels to reduce classification errors due to spurious environment correlations. Last, unlike previous approaches, we perform hyperparameters tuning of our method, without the need of any *true* group labels on the validation set.

In the following, we assume access to a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ composed of N images $\boldsymbol{x}_i$ with label $y_i \in \{1 \cdots K\}$.

Environment discovery. Previous work [21] observed that ERM tends to fit models on data presenting easy-to-learn spurious correlations at the beginning of the learning process. We hence train for a few iterations an exogeneous "identification model" –a convolutional neural network Φ composed of L layers with parameters $\boldsymbol{\omega}$, pre-trained on ImageNet [7]– by empirical loss minimization:

$$\min_{\boldsymbol{\omega}} \frac{1}{N} \sum_{i=1}^{N} \ell(\Phi(\boldsymbol{x}_i, \boldsymbol{\omega}), y_i), \tag{1}$$

where the cross-entropy loss ℓ is applied between the model's prediction $\Phi(\boldsymbol{x}_i, \boldsymbol{\omega})$ and the true label y_i for sample $\boldsymbol{x}_i$. After this initial training, and in the rest of the paper, the parameters $\boldsymbol{\omega}$ of the identification model Φ are frozen.

We now turn to feature-based clustering. We denote the feature map of an image $\boldsymbol{x}$ at layer l of Φ by $\phi_l(\boldsymbol{x}) \in \mathbb{R}^{M_l \times C_l}$, where C_l is the number of channels and M_l is the spatial dimension of the feature map. For each image $\boldsymbol{x}_i$, we extract its feature maps at $S \leqslant L$ different and fixed layers $l_1, \ldots, l_S$, and compute the Gram matrices defined as:

$$\mathsf{G}_l(\boldsymbol{x}_i) = \frac{1}{M_l}\, \phi_l(\boldsymbol{x}_i)^\intercal \phi_l(\boldsymbol{x}_i) \in \mathbb{R}^{C_l \times C_l},\ l = l_1 \cdots l_S. \tag{2}$$

We then vectorize and normalize each of these S Gram matrices:

$$\boldsymbol{f}_{i,l} = \mathrm{vec}(\mathsf{G}_l(\boldsymbol{x}_i)) / \left\|\mathrm{vec}(\mathsf{G}_l(\boldsymbol{x}_i))\right\|_2 \in \mathbb{R}^{C_l^2}. \tag{3}$$

The normalization permits us to balance evenly the contribution of each Gram matrix in the clustering loss. The "environment" of each image $\boldsymbol{x}_i$ is thus encoded by the vector $\boldsymbol{f}_i = [\boldsymbol{f}_{i,l_1}; \cdots; \boldsymbol{f}_{i,l_S}] \in \mathbb{R}^D$, where $D = \sum_{l=l_1}^{l_S} C_l^2$. We discover E' environments by clustering the N training images into E' clusters $\mathcal{C}_1, \ldots, \mathcal{C}_{E'}$, via k-means clustering, i.e., by computing a solution to:

$$\min_{\mathcal{C}_1 \cdots \mathcal{C}_{E'}} \sum_{e=1}^{E'} \frac{1}{|\mathcal{C}_e|} \sum_{i,j \in \mathcal{C}_e} \|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2, \tag{4}$$

where $\|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2 = \sum_{l=l_1}^{l_S} \|\boldsymbol{f}_{i,l} - \boldsymbol{f}_{j,l}\|_2^2$.

To overcome the computational cost of storing all these vectors and computing distances between them in high dimension, we perform random projections of the vectors $\boldsymbol{f}_{i,l}$ in a lower-dimensional space as proposed in [1] (see more details in the supplementary material)

Robust optimization with pseudo-groups labels. As a results of the clustering, each training image is now equipped with a pseudo-environment label $\hat{e} \in \{1 \cdots E'\}$. Combined with its class label, this provides a *pseudo-group label* $\hat{g} = (\hat{e}, y)$. The training set being now partitioned into pseudo-groups, we train a robust classifier h, distinct from Φ, with parameters $\boldsymbol{\theta}$, by minimizing the worst-group risk (GroupDRO [24]):

$$\min_{\boldsymbol{\theta}} \Big\{ \max_{(e,k)} \frac{1}{|\mathcal{D}_{e,k}|} \sum_{\substack{i \in \{1 \cdots N\}: \\ \hat{e}_i = e, y_i = k}} \ell\big(h(\boldsymbol{x}_i, \boldsymbol{\theta}), y_i\big) \Big\}, \qquad (5)$$

based on the cross-entropy loss ℓ, where $\mathcal{D}_{e,k} \subset \mathcal{D}$ denotes the set of samples with pseudo-group label $\hat{g} = (e, k)$.

Hyperparameters tuning without group annotation. Unlike previous approaches [2, 6, 27] that need *true* group labels in the validation set to define and assess worst-group performance as the metric to set hyperparameters, we first partition the validation set using the clusters found on the training set and then conduct cross-validation based on the resulting pseudo-groups. In our experiments, we observe that this type of model selection is effective to achieve proper group robustness.

3. Experiments

Firstly, we empirically show that GRAMCLUST outperforms, on three datasets, other baselines addressing robustness without group annotation. Secondly, we present an empirical analysis of our approach, including: the importance of using Gram matrices to capture style, the impact of the choice of layers to extract features from, and the impact of the number of clusters. The code will be published if the paper is accepted.

Datasets. We experiment with three standard image classification datasets on which previous works evaluate worst-group performance: **Waterbirds** [24] is a dataset composed of bird photographs from the CUB dataset [29] superimposed on background scenes taken from the Places365 dataset [31]. The target labels are "landbird" and "waterbird" which are spuriously correlated with the background images of either "land" or "water". We evaluate on the test set with the average accuracy and the worst-group accuracy ("waterbird" on "land"); **CelebA** [22] is a celebrity large-scale face dataset with 202,599 natural images. There exists a spurious correlation between the hair color and the gender ("male" or "female") of a person. We evaluate on the test set with the average accuracy and the worst-group accuracy ("male" with "blond" hair); **COCO-on-Places-224** is the same dataset as in [2] but with images resized to 224 × 224 (instead of 64 × 64 in the original paper). There are 10 segmented COCO [20] objects superimposed on scenes from the Places365 dataset. This time, a group of backgrounds are spuriously correlated with each object at training time. We evaluate the accuracy on a first test set with objects on the same backgrounds as during training, called the *in-distribution set* ('ind'), and on a second test set with objects on unseen backgrounds, dubbed the *systematically-shifted set* ('sys').

Baselines. We compare our approach against the standard **ERM** baseline and recent methods that aim at robust predictions across groups without the use of train group annotations (**EIIL** [6], **GEORGE** [27] and **JTT** [21]). We also include robust methods that use *true* group annotations at train time (**IRM** [3], importance weighting and **GroupDRO** [24]). The latter methods and ERM were already implemented and we took care to reproduce results for all methods. Note that our approach and GroupDRO share the same robust optimization objective.

Training details. All methods use a ResNet-50 architecture pre-trained on ImageNet [7] as the robust classifier (classifier h in Section 2). Models are optimized using SGD with momentum. For GroupDRO and ERM, we use the hyperparameters reported by the authors on Waterbirds and CelebA datasets. Note that hyperparameters have been selected with the use of a validation set with group labels. Regarding our approach, we select a VGG-19 [26] architecture for the identification model (Φ in Section 2) and train it for 1 epoch using SGD with momentum. Among usual layers used to compute style representations in neural style transfer, we observed improved performance by selecting deeper layers in the network: for each dataset, we consistently extract features from the *conv5_1* layer, i.e., the first convolutional layer of block 5. We include results with two types of model: (i) based on validation set with true-group annotations ('GRAMCLUST-*orig*'); (ii) based on pseudo-group labels ('GRAMCLUST-*cv*') predicted by our clustering.

Comparatives results. We report quantitative comparisons on Waterbirds, CelebA and COCO-on-Places-224 in Table 1. First, we observe that GRAMCLUST improves worst-group test accuracy over ERM baseline on Waterbirds and CelebA and sys accuracy on COCO-on-Places224. More importantly, GRAMCLUST-*cv* achieves state-of-the-art performances on group robustness compared to all methods that do not use group labels on the training set. On CelebA, which is a large-scale datasets with natural images, our approach outperforms the previous best, JTT [21], by

[1] Results with JTT differ from the original paper as the scores that we report correspond to models trained without early-stopping. The authors

Table 1. **Comparative results on Waterbirds, CelebA and COCO-on-Places-224.** Worst-group (*w-g*) and average (*avg*) test accuracies (% mean and std.) for Waterbirds and CelebA datasets; systematically-shifted (*sys*) and in-distribution (*ind*) test-set accuracies (% mean and std.) for COCO-on-Places dataset. Experiments with ResNet-50 models. Underlined and **bold** type indicate respectively best and per-block best performance (with significance $p < 0.05$ according to paired t-test on five runs).

	Grp labels		Waterbirds		CelebA		COCO-on-Places	
Method	train	val	w-g	avg	w-g	avg	sys	ind
ERM		✓	65.0$\pm$2.7	97.3$\pm$0.1	42.4$\pm$1.5	94.8$\pm$0.1	71.9$\pm$0.3	95.5$\pm$0.1
IRM [3]	✓	✓	77.4$\pm$0.3	**97.3**$\pm$0.1	75.1$\pm$0.6	**94.5**$\pm$0.1	78.8$\pm$0.3	**95.1**$\pm$0.2
Imp. Weighting	✓	✓	74.4$\pm$0.6	**97.4**$\pm$0.1	72.4$\pm$1.4	**94.4**$\pm$0.2	71.7$\pm$0.5	93.7$\pm$0.2
GroupDRO [24]	✓	✓	**83.9**$\pm$0.3	96.8$\pm$0.1	**85.7**$\pm$2.0	93.7$\pm$0.2	**79.0**$\pm$0.4	95.2$\pm$0.2
EIIL [6]		✓	78.7$\pm$0.3	**96.9**$\pm$0.1	-	-	68.5$\pm$0.4	94.8$\pm$0.3
GEORGE [27]		✓	76.2$\pm$2.0	95.7$\pm$0.5	53.7$\pm$1.3	**94.6**$\pm$0.2	71.6$\pm$0.3	**95.1**$\pm$0.1
JTT[1] [21]		✓	82.9$\pm$0.3	96.4$\pm$0.2	56.0$\pm$0.7	93.6$\pm$0.0	69.2$\pm$0.4	94.7$\pm$0.3
GRAMCLUST-*orig*		✓	**85.3**$\pm$1.1	96.6$\pm$0.1	**77.9**$\pm$2.2	94.2$\pm$0.2	**72.4**$\pm$0.4	**95.0**$\pm$0.2
GRAMCLUST-*cv*			**85.3**$\pm$1.1	96.6$\pm$0.1	**80.3**$\pm$1.9	93.4$\pm$0.1	**73.2**$\pm$0.3	95.3$\pm$0.3

21.9 pts. We were not able to scale EIIL [6] on this dataset due to memory overflow issues. Note that GRAMCLUST-*orig* uses the same hyperparameters as EIIL, GEORGE and JTT for robust training of the target classifier from predicted group labels, and still displays significant improvements. Surprisingly, GRAMCLUST-*cv* and GRAMCLUST-*orig* outperform GroupDRO on Waterbirds with 85.3% *vs.* 83.9%, while the latter method uses true-group labels during training. This may be due to the ambiguity of the background in some Waterbirds images.

Importance of Gram matrices. Since [15] uses the channel-wise mean and variance of image features to perform style transfer, we compare the use of such style statistics (‘*MeanVar*’) against our use of Gram matrices in Table 2. *MeanVar* reaches test worst-group accuracy on-par with *Gram matrix* on Waterbirds but degrades significantly performances on CelebA. Gram matrices provide more information than *MeanVar* as their diagonals already contain the information about the channel-wise mean and variance of the deep features (see Eq. 2). Hence, these results show that, when scaling on large and natural-image datasets such as CelebA, keeping all the correlations between different channels is important for group robustness.

Choice of layers for clustering features. We also compared our use of VGG-19 *conv5_1* features to capture style with the direct use of the penultimate (‘*AvgPool*’) representation of a more modern ResNet-50 identification model. Note that, while dating back to 2015, VGG features are still successfully used through their Gram matrices, e.g. in [5, 14, 18]. In Table 2, we observe that using the penultimate layer of a ResNet-50 as style representation for the clustering produces poorer performance.

Table 2. **Comparison of ways to capture style**. Results in worst-group (Waterbirds, CelebA) and systematically-shifted (COCO-on-Places) test-set accuracies (%). Gram matrices are more effective at capturing style toward improved group robustness.

Style feat.	Arch.	Layer	Waterbirds	CelebA	COCO-on-P
Standard	ResNet-50	*AvgPool*	76.2$\pm$2.0	53.7$\pm$1.3	71.6$\pm$0.3
MeanVar	VGG-19	*Conv5_1*	**85.3**$\pm$1.2	69.8$\pm$1.0	71.4$\pm$0.5
Gram matrix	VGG-19	*Conv5_1*	**85.3**$\pm$1.1	**77.9**$\pm$2.2	**72.4**$\pm$0.4

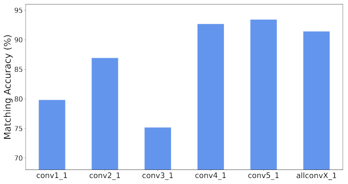

Figure 2. **Impact of the layer used by GRAMCLUST to extract style**. Group matching accuracy (with Hungarian algorithm) on the validation set on Waterbirds.

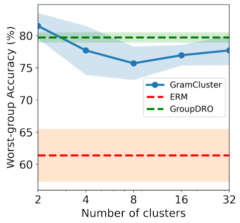

Figure 3. **Impact of cluster number.** GRAMCLUST’s Worst-group val accuracies on Waterbirds.

Clustering analysis. We study the behavior of our clustering algorithm w.r.t. the layers selected to extract features and to the number of clusters. This analysis is conducted on the Waterbirds dataset.

First, we evaluate the impact of the selection of VGG-19 layers to extract the features in the clustering stage. To this end, we study the matching of the predicted environments to the true environment labels on the validation via Hungarian matching [17] and measure the global matching accuracy across all validation samples for each five layers commonly used in neural style transfer (*conv1_1*,*conv2_1*,*conv3_1*,*conv4_1*,*conv5_1*). Results in Figure 2 show that: Features from deeper layers correlate with better matching accuracy; Our approach is robust to the choice of deep layers either taken together (*allconvX_1*) or individually such as *conv4_1* and *conv5_1*; Using *conv5_1* outperforms selecting all traditional style layers. Consistent conclusions are found on the CelebA dataset (see Supplementary material).

Second, we study the impact of the number of clusters as hyperparameter in the clustering algorithm. Worst-group accuracy on the validation set for $E' \in \{2, 4, 8, 16, 32\}$ clusters are reported in Figure 3. Or method is robust to a variation in the number of clusters: GRAMCLUST with more clusters than actual environments produces a slight drop in performance but still improves performance over

select models before convergence (around epoch 3) with low average accuracy on the test set but high worst-group accuracy

ERM and remains on-par performances with GroupDRO.

4. Conclusion

In this paper, we introduce GRAMCLUST, a two-stage method that first partitions a training dataset into style-based clusters via k-means algorithm based on Gram matrices computed from features, themselves extracted from an identification model trained to catch spurious correlations of a biased dataset. This first stage is then followed by learning a robust classifier by minimizing the error on the worst pseudo-group labels previously discovered. GRAMCLUST demonstrates to be an effective approach to tackle group robustness and outperforms every single baseline on standard datasets with spurious correlations. The usage of feature Gram matrices is of primary importance to correctly characterize the environment of the image and enables a relevant partition for robust training. Our approach also alleviates the need to label a small validation set of images with group information and is able to tune its hyperparameters without group supervision by applying its clustering algorithm on the validation set.

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, pages 671—-687, 2003. 2, 7

[2] Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 3, 7

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. 1, 3, 4, 7

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[5] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 872–881, 2021. 4

[6] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 1, 2, 3, 4, 7, 8

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2, 3

[8] John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2, 2019. 1

[9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 8

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[11] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7, 8

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[14] Lukas Höllein, Justin Johnson, and Matthias Nießner. Stylemesh: Style transfer for indoor 3d scene reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6198–6208, 2022. 4

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

[16] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 2, 7

[17] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955. 4

[18] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 4

[19] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *International Joint Conference on Artificial Intelligence*, page 2230–2236, 2017. 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. 3

[21] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 1, 2, 3, 4, 7

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 3

[23] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 1

[24] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 3, 4, 7

[25] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020. 1

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 3, 8

[27] Nimit Sharad Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. 1, 2, 3, 4, 7, 8

[28] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59, 2017. 1

[29] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 3

[30] Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations*, 2021. 1

[31] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(6):1452–1464, 2018. 3

Appendices

A. Details on random projection

Storing all flattened Gram matrices and computing distances between them in a high-dimensional space is computationally and memory expensive on large datasets. We overcome this difficulty by projecting the vectors $\boldsymbol{f}_{i,l}$ (Eq. 3) in a lower-dimensional space as proposed in [1]. We build a matrix $\mathsf{P} \in \mathbb{R}^{\ell_0 \times D}$ whose entries P_{mn} are the realisation of independent random variables: $\mathsf{P}_{mn} = 1$ or $\mathsf{P}_{mn} = -1$ with probability $1/2$. Then we compute

$$\tilde{\boldsymbol{f}}_{i,l} = \frac{1}{\sqrt{\ell_0}} \mathsf{P} \boldsymbol{f}_{i,l} \tag{6}$$

and substitute $\tilde{\boldsymbol{f}}_{i,l}$ for $\boldsymbol{f}_{i,l}$ in Eq. 4. We justify this choice by the fact that this projection preserves the distances $\|\boldsymbol{f}_{i,l} - \boldsymbol{f}_{j,l}\|_2^2$ involved in the k-means objective. Indeed, let $\varepsilon \in]0, 1[$ and $\ell_0 \propto \log(N)$, then with high probability:[2]

$$(1-\varepsilon) \|\boldsymbol{f}_{i,l} - \boldsymbol{f}_{j,l}\|_2 \leqslant \|\tilde{\boldsymbol{f}}_{i,l} - \tilde{\boldsymbol{f}}_{j,l}\|_2 \leqslant (1+\varepsilon) \|\boldsymbol{f}_{i,l} - \boldsymbol{f}_{j,l}\|_2 \,, \tag{7}$$

for all i and j in $\{1 \cdots N\}$. In practice, we choose $\ell_0 = \lfloor 100 \log(N) \rfloor$ which yields dimensions for $\tilde{\boldsymbol{f}}_{i,l}$ much lower than typical values of D. We remark that this choice of projection is independent of all $\boldsymbol{f}_{i,l}$ and thus can be defined and fixed before any feature extraction.

B. Implementation details

This section focuses on implementation details used to produce the results in the main text of our paper. Our implementation builds upon the WILDS framework[3] released with the paper of Koh *et al.* [16].

B.1. Construction of COCO-on-Places-224

We generated the dataset using the code[4] of Ahmed *et al.* [2] but, as explained in the main paper, we modified it to produce images of size 224 × 224 instead of 64 × 64. The reader can refer to the appendix of [2] for more details regarding the generation of the COCO-on-Places dataset.

B.2. Details about robust optimization

We trained all models on one NVIDIA® V100 Tensor Core with 16GB of memory, using PyTorch 1.10 and CUDA 10.2.

We used the implementations of IRM [3], Importance Weighting and GroupDRO [24] available in WILDS [16], our own implementations of JTT [21] and of GEORGE [27] (while making sure that we could reproduce the original performances on Waterbirds and CelebA), and the official implementation[5] of EIIL [6]. Concerning EIIL, we recall that we were not able to make this method scale to large datasets such as CelebA.

For all methods, we used a ResNet-50 [12] architecture trained using stochastic gradient descent with momentum (SGD-M) and L_2 regularization, but without any learning rate scheduler. We used a momentum of 0.9 and a batch size of 128 for all datasets and all methods. The learning rate η and L_2 regularization parameters λ are set as detailed below.

JTT, GEORGE, EIIL, GRAMSTYLE all use GroupDRO [24] as robust optimization step. On Waterbirds and CelebA, we did not redo any grid search and used the hyperparameters found in [24]. These hyperparameters were optimized using a small validation set annotated with true group labels. To produce the results on COCO-on-Places-224, we performed our own grid search using the annotated validation set. We considered values of η and λ close to those used in [24]: $\lambda \in \{10^{-4}, 10^{-2}, 10^{-1}, 1\}$ and $\eta \in \{10^{-5}, 5 \cdot 10^{-5}, 10^{-4}\}$. The best hyperparameters for GroupDRO are summarized in Table 3.

To ensure fair comparisons, we also performed the same grid search over η and λ for ERM, IRM and Importance Weighting. The best hyperparameters for ERM and IRM are summarized for each dataset in Tables 4 and 5, respectively. Note that they correspond to those reported in [24] for Waterbirds and CelebA.

Table 3. **SGD-M hyperparameters for GroupDRO training**.

SGD-M hyperparam.	Waterbirds	CelebA	COCO-on-P
Learning rate η	10^{-5}	10^{-5}	$5 \cdot 10^{-5}$
L_2 regularization λ	1.0	0.1	10^{-2}

Table 4. **SGD-M hyperparameters for ERM training**.

SGD-M hyperparam.	Waterbirds	CelebA	COCO-on-P
Learning rate η	10^{-4}	10^{-4}	10^{-4}
L_2 regularization λ	10^{-3}	10^{-4}	10^{-4}

Table 5. **SGD-M hyperparameters for IRM training**.

SGD-M hyperparam.	Waterbirds	CelebA	COCO-on-P
Learning rate η	10^{-4}	10^{-5}	$5 \cdot 10^{-5}$
L_2 regularization λ	10^{-3}	0.1	0.1

[2] We let the reader refer to Theorem 1.1 in [1] for the exact expression of this probability as a function of ε, N and ℓ_0.

[3] https://github.com/p-lambda/wilds

[4] https://github.com/Faruk-Ahmed/predictive_group_invariance

[5] https://github.com/ecreager/eiil

Table 6. **Grid search for GramStyle-*cv*'s hyperparameters on validation sets of Waterbirds, CelebA and COCO-on-Places-224 with pseudo-group labels.** We report the worst-group ('w-g') and average ('avg') accuracies for Waterbirds and CelebA datasets, and the systematically-shifted ('sys') and in-distribution ('ind') accuracies for COCO-on-Places dataset.

Hyperparam.		Waterbirds		CelebA		COCO-on-P	
λ	η	w-g	avg	w-g	avg	sys	ind
0.01	$1 \cdot 10^{-5}$	74.6	82.4	**86.0**	93.2	62.8	92.3
0.01	$5 \cdot 10^{-5}$	69.2	79.9	53.5	**94.6**	70.7	76.5
0.01	$1 \cdot 10^{-4}$	70.0	80.6	-	-	78.5	82.7
0.1	$1 \cdot 10^{-5}$	75.4	82.6	85.6	93.7	**78.7**	83.3
0.1	$5 \cdot 10^{-5}$	73.8	82.4	85.0	89.1	70.4	76.4
0.1	$1 \cdot 10^{-4}$	76.9	85.8	-	-	76.2	81.2
1	$1 \cdot 10^{-5}$	**80.8**	**86.4**	-	-	65.5	72.6
1	$5 \cdot 10^{-5}$	0.0	23.1	-	-	0.1	11.1
1	$1 \cdot 10^{-4}$	0.0	23.1	-	-	0.2	11.1

B.3. Group discovery details

For GramStyle, we follow the standard practice of neural style transfer [9] and use the VGG-19 [26] architecture for the identification model. This model is trained during 1 epoch on the training dataset with ERM using a batch size of 128 and SGD-M. In the experiments of Section 4.2 in the main paper, we set the number of clusters to 2, and use the layer *conv5_1* to extract Gram Matrices.

For EIIL and GEORGE, the identification model is a ResNet-50 [12] as used in the original methods. We train the model for 1 epoch with ERM using SGD-M, as for GramStyle. Note that the activation at the output of the last layer is a sigmoid in EIIL [6] while it is a softmax in GEORGE [27]. As for GramStyle, the best results were obtained when using 2 clusters for EIIL and GEORGE. We refer the reader to [6] and [27] for other implementation details specific to EIIL and GEORGE, respectively.

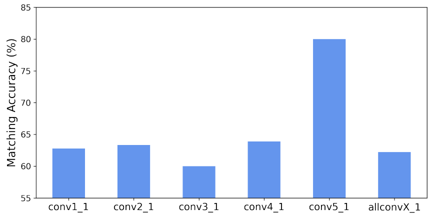

Figure 4. **Impact of the layer choice to extract style on CelebA**. We show the matching accuracy between the ground-truth environments on the validation set of CelebA and the discovered ones with GramStyle when using different VGG-19 layers. The result denoted *allconvX_1* is obtained when using all the layers *conv1_1*, *conv2_1*, *conv3_1*, *conv4_1*, *conv5_1* in our method.

B.4. Cross validation on pseudo-group annotations

We report in Table 6 the results of our grid search on the validation set of each dataset using the *pseudo-annotations* discovered with our method, i.e., using our discovered environments instead of the ground-truth ones. Hence, the average and worst-group accuracies in Table 6 are computed using the discovered pseudo-groups. The hyperparameters used in GramStyle-*cv* correspond to those that yield the best worst-group accuracy in this table.

C. Clustering analysis on CelebA

We present, in Figure 4, the matching accuracy between the ground-truth environments and the environments discovered with our method on the validation set of CelebA for different layers of the VGG-19. As on Waterbirds, we notice that the best result is obtained when using the layer *conv5_1*.