Unsupervised Concept Discovery Mitigates Spurious Correlations

Md Rifat Arefin [1] **Yan Zhang** [2] **Aristide Baratin** [2] **Francesco Locatello** [3] **Irina Rish** [1] **Dianbo Liu** [4]
Kenji Kawaguchi [4]

Abstract

Models prone to spurious correlations in training data often produce brittle predictions and introduce unintended biases. Addressing this challenge typically involves methods relying on prior knowledge and group annotation to remove spurious correlations, which may not be readily available in many applications. In this paper, we establish a novel connection between unsupervised object-centric learning and mitigation of spurious correlations. Instead of directly inferring subgroups with varying correlations with labels, our approach focuses on discovering *concepts*: discrete ideas that are shared across input samples. Leveraging existing object-centric representation learning, we introduce CoBalT: a concept balancing technique that effectively mitigates spurious correlations without requiring human labeling of subgroups. Evaluation across the benchmark datasets for sub-population shifts demonstrate superior or competitive performance compared state-of-the-art baselines, without the need for group annotation. Code is available at https://github.com/rarefin/CoBalT

1. Introduction

A critical concern with deep learning models arises from their well-known tendency to base their predictions on correlations present in the training data rather than robustly informative features (Arjovsky et al., 2019; Sagawa et al., 2020). For instance, in image classification, translating an image by a few pixels (Azulay & Weiss, 2019) or modifying the background (Beery et al., 2018) can drastically change the predictions of the model. Often viewed as resulting from the so-called 'simplicity bias' of deep neural networks in the literature (Shah et al., 2020), this phenomenon pervades the landscape of deep learning models (Geirhos et al., 2020).

While models relying on spurious correlations may perform well on average across i.i.d. test data, they often struggle on specific subgroups where these correlations do not hold. Common approaches involve partitioning the training data based on prior knowledge of spurious information and adjusting the training process to ensure consistency across these groups (Sagawa et al., 2020; Kirichenko et al., 2023; Arjovsky et al., 2019). However, most real-world datasets lack explicit annotations highlighting spurious information. Manual annotation is expensive and can be ill-defined, as the appropriate groupings may not be immediately apparent.

On the other hand, self-supervised learning (Chen et al., 2020; Caron et al., 2020; 2021; Grill et al., 2020; He et al., 2020) has produced powerful representation learners. Several methods (Cho et al., 2021; Wen et al., 2022) aim to learn high-level concepts by semantic grouping of areas within an input image into object-centric instances. Wen et al. (2022), for instance, leverage slot attention (Locatello et al., 2020) to decompose complex scenes into constituent objects via contrastive learning alone. While their original aim was downstream task representation learning, we posit that such decomposition can help mitigating spurious correlations. By treating semantic groupings as concept sources discovered by the model, they can serve as data-driven proxies of subgroup labels. This differs from existing work in spurious correlation, which typically directly infers subgroups (see related work in Section 2). Our approach models concepts that do not necessarily correspond directly to subgroups; typically, we use a significantly larger number of concepts than annotated subgroups in the dataset.

This paper demonstrates the use of object-centric representation learning approaches to design classifiers robust to spurious correlations without the need for human-labeled subgroup annotations. We introduce CoBalT, a method combining concept discovery with concept balancing for robust classification. CoBalT follows a two-stage procedure common in the literature: first, inferring information about the training data, and then leveraging this information for robust training.

In **Stage 1**, we propose to vector quantize semantic group-

[1]Mila, University of Montreal, Canada [2]Samsung - SAIT AI Lab, Montreal, Canada [3]Institute of Science and Technology Austria [4]National University of Singapore. Correspondence to: Md Rifat Arefin <rifat.arefin@mila.quebec>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

ing representations into discrete concepts (Section 3.2), enabling the association of each input with relevant sets of concepts (see Fig 1) and facilitating the calculation of concept occurrence statistics across the dataset.

In **Stage 2**, we utilize the occurrence statistics of concepts via importance sampling to train a separate classifier (Section 3.3). The architecture of the classifier is inconsequential; the key contribution lies in the concept-aware sampling procedure, bridging object-centric representation learning and learning under subpopulation shifts.

Integrating Stages 1 and 2 introduces CoBalT (Concept Balancing Technique) tailored for robust classification. We evaluate CoBalT across the CMNIST, Waterbirds, CelebA, Urban Cars and ImageNet-9 (IN-9L) datasets, demonstrating improvements without the need for group annotations (Section 4). We achieve a 3% improvement on Waterbirds compared to state-of-the-art (SOTA) group agnostic methods like MaskTune (Asgari et al., 2022), ULA (Tsirigotis et al., 2023) and XRM (Pezeshki et al., 2023), remain competitive on CelebA, and achieve 1–2% improvement on challenging IN-9L test sets while maintaining original test set performance. We also show 3% improvement over SOTA baseline requiring group annotation in the Urban Cars dataset containing multiple spurious correlations per class.

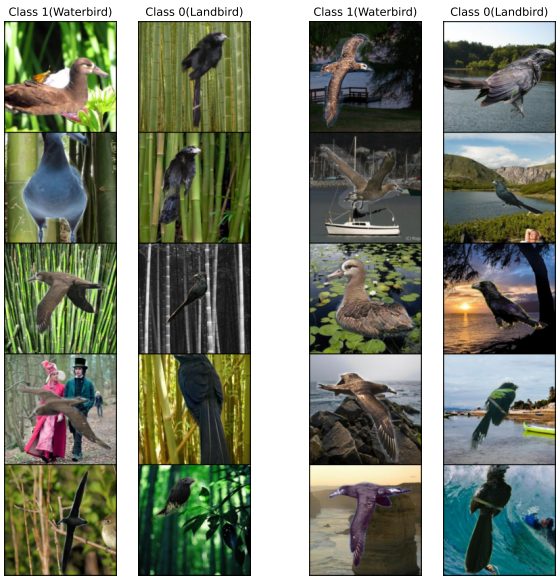

Figure 1. Images from Waterbirds dataset with different discovered concepts through our method. Here we arbitrarily select two of these concepts, which can be interpreted as trees/bamboo background (left) and water background (right), and show input samples from each of these.

2. Related Works

Robust training: Our approach extends existing methodologies for robust classification model training, particularly addressing the costly acquisition of group labels in real-world data. Unlike GDRO (Sagawa et al., 2020), which optimizes for the worst group-level error, and its semi-supervised extension, SSA (Nam et al., 2022), our method is tailored for scenarios lacking sufficient labeled group data. Additionally, methods like DFR (Kirichenko et al., 2023) and AFR (Qiu et al., 2023) retrain the classification layer with group-balanced datasets and ensure feature reweighing, requiring group-labeled training/validation data, a necessity we circumvent. ULA (Tsirigotis et al., 2023) employs a Self-Supervised Learning (SSL) pre-trained model's predictions as a bias proxy, while MaskTune (Asgari et al., 2022) assumes predictions from Empirical Risk Minimization (ERM) models to be biased. To train an unbiased model, the former adjusts the classifier's logits during debiasing training, and the latter masks out the input data based on the saliency map of the prediction.

Group inference methods: Obtaining group labels in real-world data is often costly. Several methods have been proposed for inferring group labels initially, followed by robust model training. LfF (Nam et al., 2020) uses two models, where the second model is trained using examples with higher loss in the first model. This approach contrasts with GEORGE (Sohoni et al., 2020), which clusters representations from the first stage ERM model to infer group information and then trains a second model using GDRO. Similarly, JTT (Liu et al., 2021) and and CNC (Zhang et al., 2022)identify minority groups based on miss-classifications of the first stage ERM model; however, JTT continues with ERM to train the robust model, while CNC uses contrastive learning to align representations of minority examples with the majority. These methods either rely on extra group annotation or fail in the presence of multiple unbalanced minority groups and noisy examples (Yang et al., 2024).

SPARE (Yang et al., 2024) separates spurious information in the early stages of training and uses k-means clustering to differentiate between minority and majority groups but relies on validation group annotation data to determine the specific epoch for separation. Conversely, our approach does not depend on group-annotated data for epoch identification; instead, we utilize a self-supervised method combined with spatial decomposition to separate spurious and non-spurious information effectively.The recently introduced XRM (Pezeshki et al., 2023) identifies groups within training and validation datasets through model prediction errors, operating under the assumption that models inherently learn spurious correlations. This methodology could be detrimental in scenarios where such an assumption does not hold true (Yong et al., 2022).

Concept discovery. To address spurious correlations, DISC (Wu et al., 2023) introduces a human-interpretable concepts bank and Moayeri et al. (2023) rank data points by the spu-

rious concepts that they contain. However, they require additional annotations of potential spurious features, posing practical challenges and limiting its general applicability. Learning abstract representations from images by decomposing them into higher-level concepts without human annotations has been explored in previous work. A recent development is slot attention (Locatello et al., 2020), which groups spatially repetitive visual features by imposing an attention bottleneck. This method and its variants have been successfully applied to discover object-centric concepts in synthetic datasets (Locatello et al., 2020; Engelcke et al., 2021; Zhang et al., 2023). However, they face challenges when applied to complex real-world data. Seitzer et al. (2023) hypothesized that reconstructing the pixel space as a learning objective might not introduce enough inductive bias to facilitate the emergence of objects or concepts in real data. As a solution, they propose reconstructing the features from the self-supervised pre-trained DINO model (Caron et al., 2021). With similar motivation, Wen et al. (2022) employs a joint embedding teacher-student architecture, similar to Caron et al. (2021), where the student model attempts to predict the concept representations of the teacher network. We extend this work to discover discrete symbols-like concepts by applying vector quantization (van den Oord et al., 2017) to continuous concept representations aiding compositional reasoning of images, such as identifying common groups or attributes in the dataset like humans.

3. Method

Unlike existing methods, our goal is *not* to discover the subgroups of a dataset specifically, but more general *concepts*. For example, while groups in the Waterbirds dataset are explicitly defined to be the product of classes with some binary background attribute, $\{\text{water bird}, \text{land bird}\} \times \{\text{water background}, \text{land background}\}$, the concepts could capture dataset-independent ideas such as blue bird, street background, or short beak.

We base our approach on the two-stage training procedure common in the literature, with the first stage determining some information about the training data and the second stage using this information to perform robust training.

The first stage combines two key components:

1. Spatial clustering (Caron et al., 2021; Wen et al., 2022), which groups pixels into semantic regions (Section 3.1). While our approach in this paper is based on the method by Wen et al. (2022), in principle, the requirement is simply for an unsupervised representation learner that decomposes the input into objects.
2. A novel vector clustering technique we call *concept dictionary learning* (Section 3.2), achieved through vector quantization (van den Oord et al., 2017). This process discretizes the information of the slots into distinct concepts, which are more manageable compared to continuous representations of semantic regions. For example, instead of storing details about the specific shape and appearance of a bird, this clustering identifies broader concepts like bird types, which offer greater utility across various inputs. These concepts encompass typical foreground objects such as cats and dogs, background elements like land and sky (see Fig. 1), or other abstract notions not as readily interpretable as individual words.

The key aspect of our proposed second component is its independence from human labeling, achieved through leveraging the self-supervised learning setup of the first component. This lack of reliance to human labeling offers significant advantages, particularly in complex datasets. For example, when dealing with large datasets like ImageNet, determining relevant subgroups across the images is challenging due to the vast number of possibilities. Spurious correlations in a dataset are likely to vary depending on the specific task being performed with the dataset. Without pre-labeling every conceivable group (which is clearly infeasible), identifying the subgroups necessary to address spurious correlations seems nearly impossible.

By adopting a data-driven approach where concepts are learned, we can discover concepts that a model inherently relies on. However, this approach has the potential drawback of weakening the connection between a learned concept and a concept that humans readily understand. One advantage of an object-centric decomposition, as demonstrated by methods like Wen et al. (2022), is that the spatial grouping of a concept provides humans with additional insight into that concept represents.

3.1. Architecture

The model architecture used for concept learning shares the same overall structure as many recent self-supervised approaches to representation learning (Caron et al., 2021; Grill et al., 2020; Zbontar et al., 2021; Chen et al., 2020). Following Wen et al. (2022), we employ a two-branch network where the branches are structurally similar but asymmetric in parameter weights. Each branch comprises an *encoder* that outputs patch representation vectors of the input image, a *projector* that transforms the representations into an embedding space, and a *slot module* where spatial patch representation vectors are semantically grouped into concept representations. Our focus lies on building our model based on the output of the slot module. The overall architecture is illustrated and briefly described in Figure 2, with detailed information provided in Appendix A.

More precisely, we will utilize the slots of the student and teacher branches, $z_s \in R^{N \times d}$ and $z_t \in R^{N \times d}$, where the

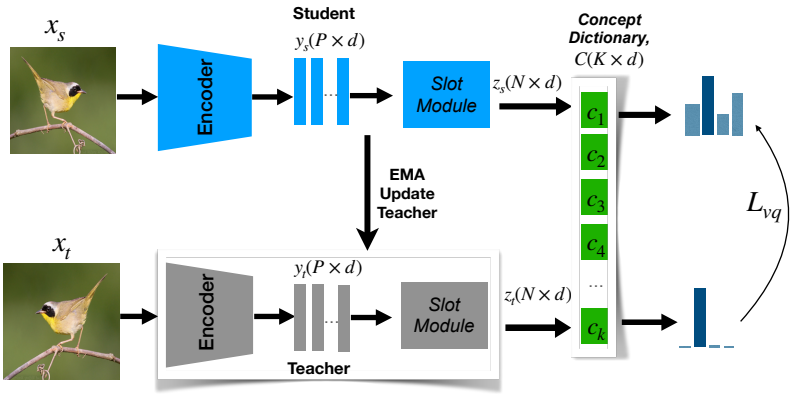

Figure 2. Architecture for learning slots and clustering without human annotation. x_s and x_t are two different augmented views of the same image. The teacher and student encoders project the augmented images into different patches y_t and y_s respectively, which are subsequently decomposed into concept representations z_t and z_s by slot attention. Then z_t and z_s are clustered into different concepts of dictionary C using a vector-quantization. Stop-gradient is applied to teacher branch and the teacher encoder and slot module parameters are updated through the exponential moving average of the student encoder and slot module parameters.

hyperparameters N represents the number of slots and d denotes the dimensionality of each slot. Each slot serves as a semantic grouping of an area in the input; for instance, a slot could correspond to a single object in the image.

3.2. Concept Dictionary

For the next step, we aim to discover meaningful discrete concepts from these spatially-decomposed semantic groupings. To do so, we employ vector quantization (van den Oord et al., 2017) which acts as a learned discretization or clustering mechanism that effectively clusters similar concepts in the training data into distinct categorical concepts.

This is done by utilizing a codebook that we call **concept dictionary** $C \in R^{K \times d}$ with K vectors of dimension d, each of which corresponds to a symbolic concept (e.g. water, tree, bird, etc.) that we want to learn. Note that we do not supervise these concepts in any way – these words simply denote possible meanings we could assign to these concepts post-hoc. We assign each slot (vector representation) to a discrete symbolic concept by learning a categorical distribution over the entries in the dictionary.

Given a randomly initialized concept dictionary C, we associate each slot (for student and teacher branches) to a concept in the concept dictionary through distributions $p_s \in \mathbb{R}^{K \times N}, p_t \in \mathbb{R}^N$ by seeking the most similar concept:

$$(p_s)_{ij} = \frac{\exp(-\|C_i - (\bar{z}_s)_j\|_2^2/\tau_s)}{\sum_{t=1}^{K} \exp(-\|C_t - (\bar{z}_s)_j\|_2^2/\tau_s)} \quad (1)$$

$$(p_t)_i = \arg\max_j -\|C_j - (\bar{z}_t)_i\|_2^2 \quad (2)$$

where $C_i, (\bar{z}_s)_j \in \mathbb{R}^d$, $\bar{z}_s = z_s/||z_s||$, $\bar{z}_t = z_t/||z_t||$ (the i-th slot of z_s and z_t are normalized to have unit L2 norm), and τ_s is a temperature hyperparameter. For the teacher branch, inspired by Caron et al. (2021), rather than taking a $\mathrm{soft\,max}$, we use a sharpened distribution. In particular, we employ the $\arg\max$ to facilitate a hard assignment into a one-hot representation. This hard assignment ensures that each slot is associated with a single distinct concept. We then use these as the supervision signal for the student branch, where we encourage each slot representation of the student to also be assigned to a single concept vector. This difference in $\mathrm{soft\,max}$ and $\arg\max$ has the benefit of making the distributions of the student and teacher branches different, which avoids the representation collapse problem mentioned in Caron et al. (2021).

Following Roy et al. (2018), throughout training at each step, C is updated by the exponential moving average of batch-wise teacher concept representations z_t as follows:

$$C_j = \alpha_c \cdot C_j + (1 - \alpha_c) \cdot \sum_i \mathbb{1}\{(p_t)_i = j\}(z_t)_i \quad (3)$$

with α_c is the update rate of the codebook. We set it to 0.9 for all our experiments.

Loss As for the learning objectives, in addition to the losses proposed by Wen et al. (2022) (see Appendix A), we include a novel term $\mathcal{L}_{vq}$, motivated as follows. Since we do not have any explicit human supervision of concepts, we exploit the assignment of concepts of the teacher to supervise the student. The purpose of the loss term $\mathcal{L}_{vq}$ is to ensure the consistency of the prediction between the slot representations of the teacher and the student. We encourage this alignment by distilling the teacher's prediction of discrete concepts to the student with a cross-entropy loss, which is

calculated as follows:

$$\mathcal{L}_{vq} = -\sum_{i=1}^{N}\sum_{j=1}^{K} \mathbf{I}(i)\mathbb{1}\{(p_t)_i = j\}\log(p_s)_{ij} \quad (4)$$

where I(i) is the indicator function that avoids calculating the loss for the slot where the student and teacher does not have any common patch assignment. Details of how this is calculated are described in Appendix A.

We then include this objective as a term in the overall loss of Wen et al. (2022):

$$\mathcal{L} = \mathcal{L}_{dis} + \mathcal{L}_{con} + \mathcal{L}_{vq} \quad (5)$$

where $\mathcal{L}_{dis}$ governs *attention distillation* from teacher to student and $\mathcal{L}_{con}$ is a *contrastive loss* between slot representations to avoid redundancy and encourage diversity. These losses are described in detail in Appendix A.

This concludes the first stage of our training process. To recap, we extract slot representations following the methodology in Wen et al. (2022), then compute concept distributions p_s and p_t over the concept dictionary C, which is incrementally updated based on assignments from the teacher branch. Our learning objective is designed to distill the teacher concept distribution to the student. Through this process, we establish the association of training samples with sets of concepts. This information will be utilized in the subsequent section.

3.3. Training a Robust Classifier

In the second stage, we train a separate classifier based on the concepts learned in the first stage, which are considered fixed. Integrating this information into the training process offers various possibilities. Our approach draws inspiration from previous works (Sagawa et al., 2020; Yang et al., 2024) where, if ground-truth subgroups are known, adjusting the subgroup sampling rate evenly is the most effective method. We adapt this concept to our framework, modifying it to suit learned concepts rather than ground-truth subgroups. However, this adaptation presents challenges, such as each data point belonging to multiple concepts instead of a single subgroup, and the occurrence of each concept in multiple classes at varying frequencies.

Sampling method Our core approach involves adjusting the sampling rate of samples to ensure an even representation of concepts and, when feasible, an even representation of classes within those concepts. This entails sampling prevalent concepts less frequently and rare concepts more frequently. Additionally, within each concept, we aim to maintain balanced representation of labeled classes. By doing so, we bias the classifier training towards rarer concepts while striving to balance classes within a concept whenever possible.

This strategy is guided by the understanding that minority groups, characterized by rarer concepts within a class, are more susceptible to misclassification due to concept overlap. Notably, our sampling method differs from the weighting scheme proposed by (Yang et al., 2024), which contrasts between groups within the same class. Instead, our approach focuses on contrasting between samples from the same concept but belonging to different classes.

Within a cluster c, we have multiple classes with $T_{c,y}$ represents the samples from the cluster c and class y. We compute the weight and probability of sampling that class within the cluster as:

$$w_{c,y} = \frac{1}{|T_{c,y}|}, \quad p_{c,y} = \frac{w_{c,y}^{\lambda}}{\sum_{\hat{y}} w_{c,\hat{y}}^{\lambda}} \quad (6)$$

where λ is a sampling factor, a hyperparameter. Yang et al. (2024) recommend to increase λ from the default of 1 when the inter-concept groups are not well separable. The choice of this hyperparameter can be guided by the average silhouette score (Rousseeuw, 1987), which measures how well the clusters are separated. In our case, it reflects the degree of distinction between groups from one cluster to the groups of the other cluster.

Algorithm 1 Batch Sampling Strategy

K: clusters with samples of different classes
n: batch size
$T_{c,y}$: Set of samples belonging to cluster c and class y

Initialize $batch \leftarrow \{\}$
for $i = 1$ to n **do**
  $c \leftarrow$ uniformly select a cluster from 1 to K
  $w_{c,y} \leftarrow$ calculate weights $\frac{1}{|T_{c,y}|}$
  $y \leftarrow$ select a class with $p_{c,y} = \frac{w_{c,y}^{\lambda}}{\sum_{\hat{y}} w_{c,\hat{y}}^{\lambda}}$
  $b \leftarrow$ select a sample from c of class y
  $batch \leftarrow batch \cup \{b\}$
end for
return $batch$

3.4. Early stopping

As demonstrated by Idrissi et al. (2022), having access to group information is crucial for effective model selection, particularly in scenarios involving spurious correlations. In our experiments, we explore three distinct model selection strategies by altering the criteria for early stopping:

1. CoBalT$_{hg}$: This strategy relies on human-annotated worst-group labels.
2. CoBalT$_{ig}$: Here, we utilize the inferred worst group.
3. CoBalT$_{avg}$: This strategy employs the average validation accuracy as the criterion for early stopping.

While CoBalT$_{hg}$ offers the advantage of leveraging human annotations, it also reintroduces dependency on manual labeling. Consequently, we generally prefer settings where CoBalT$_{ig}$ and CoBalT$_{avg}$ are more suitable.

For CoBalT$_{ig}$, our approach involves inferring groups from the discovered concepts. Each group is defined by the unique combination of class and concept. For instance, if we have two concepts and three classes, we would generate six groups accordingly. It is important to note that these inferred groups may not align with the ground-truth groups in the dataset, if such labels are even available. Nevertheless, our method utilizes these inferred groups as an early stopping criterion.

4. Experiments

To illustrate the effectiveness of our spatial concept discovery and sampling strategy, we investigate two challenging scenarios where training a robust classifier using empirical risk minimization (ERM) with i.i.d. (independent and identically distributed) sampling faces significant difficulties.

Scenario 1: Classification complicated by class imbalance and attribute imbalance with single spurious correlation. In this scenario, markedly underrepresented, and attributes within classes exhibit uneven distributions in the training data. This presents considerable challenges for an ERM-trained model, particularly concerning under-represented attributes.

Scenario 2: When data containing two spurious correlations per class, Li et al. (2023) shows the limitations of SOTA methods that present 'whack-a-mole' behavior: mitigating one spurious correlation but amplifying the other.

Scenario 3: Test data has attributes not present in the training data, requiring attribute generalization. For example, if the training set has cows on grassland and rarely on a beach, the test set might have cows on a volcano. This scenario is demanding as the classifier must recognize and generalize unknown attributes. Moreover, merely defining subgroups in this scenario is inherently challenging.

4.1. Datasets

Considering the scenarios outlined, we train our model using the following publicly available datasets: CMNIST (Alain et al., 2015), CelebA (Liu et al., 2014), Waterbirds (Sagawa et al., 2020), UrbanCars (Li et al., 2023), Background Challenge ImageNet-9 (Xiao et al., 2021).

Scenario 1. The **CMNIST** dataset (Alain et al., 2015) contains colored versions of the MNIST digits (LeCun et al., 1998). We use the challenging 5-class setup from (Zhang et al., 2022), where each class pairs two digits, with 99.5% of training samples in each class spuriously correlated to a unique color.

The **CelebA** dataset (Liu et al., 2015) shows a significant class imbalance in gender (male/female) and hair color (dark/blonde). Most of the male images (162,770) have dark hair, while only 1,387 (0.85%) have blonde hair. This imbalance risks bias, potentially causing the model to associate gender with hair color.

The **Waterbirds** dataset, as detailed in (Sagawa et al., 2020), has two classes: landbirds and waterbirds. The background: land or water acts as a spurious attribute. The common instances (waterbird, water) and (landbird, land) make it challenging to differentiate the bird type from the spuriously correlated background.

Scenario 2. The **UrbanCars** dataset (Li et al., 2023) includes two classes (urban and country vehicles) along with two incidental spurious attributes: (1) background (BG): city vs. countryside and (2) co-occurring objects (CoObj): fireplug and stop sign vs. cows and horses.

Scenario 3. We utilize the **Background Challenge ImageNet-9 (IN-9L)** dataset (Xiao et al., 2021), derived from a subset of ImageNet known as ImageNet-9. This dataset is purposefully crafted to assess the robustness of models against background variations. It encompasses four distinct types of background modifications in its test sets:

- **Original:** Maintains the original background.
- **Mixed-same:** Replaces the background taken from another image within the same class.
- **Mixed-rand:** Replaces the background taken from a randomly selected image.
- **Only-FG:** Eliminates the background entirely, leaving only the foreground object.

This dataset challenges classifiers to remain robust to background changes, serving as a benchmark for evaluating a model's ability to generalize and focus on primary object features despite background variability or absence.

4.2. Results

We present additional results and ablations in Appendix C.

4.2.1. Scenario 1 (CMNIST, Waterbirds, CelebA)

As shown in Table 1, our evaluation on Waterbirds, CelebA and CMNIST showcases the effectiveness of our approach, which achieves superior or comparable performance compared to methods that do not rely on human-annotated group labels. Particularly noteworthy is CoBalT$_{ig}$, which outperforms in worst-group accuracy the recent methods ULA (Tsirigotis et al., 2023) and XRM (Pezeshki et al., 2023) by

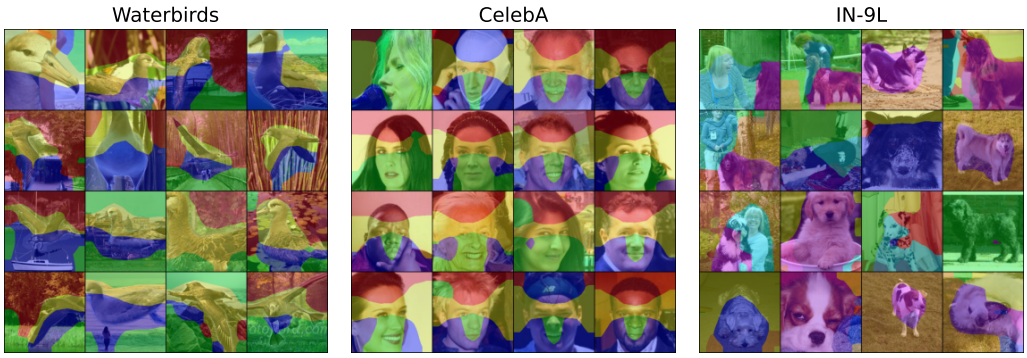

Figure 3. Segmented regions with slots in different datasets. # of slots used for Waterbirds, CelebA and ImageNet-9 is 4. The pixels in the images are grouped by slots and represent high-level concepts such as body parts of birds and backgrounds like trees, water, and so on in Waterbirds; humans, animals, grasses, and so on in *IN-9L* and nose, head, and so on in CelebA.

Table 1. CMNIST results with LeNet-5, and Waterbirds and CelebA results with ImageNet pre-trained ResNet50. Other model results are reported from Yang et al. (2024); Asgari et al. (2022); Pezeshki et al. (2023); Tsirigotis et al. (2023). The upper section uses human annotated group information for training and validation, the middle for validation only, and the bottom does not use group annotation. Best results are highlighted in each section. CoBalT$_{hg}$, CoBalT$_{ig}$, CoBalT$_{avg}$ are our trained models where early stopping is done by human-annotated worst group, inferred worst group, and average validation accuracy respectively.

	Group Label		CMNIST		Waterbirds		CelebA	
Method	**Train**	**Val**	Worst Group	Average	Worst Group	Average	Worst Group	Average
GB	✓	✓	**82.2±1.0**	**91.7±0.6**	86.3±0.3	93.0±1.5	85.0±1.1	92.7±0.1
DFR^{Tr} (Kirichenko et al., 2023)	✓	✓	-	-	90.4±1.5	**94.1±0.5**	80.1±1.1	89.7±0.4
DFR^{Val} (Kirichenko et al., 2023)	✓	✓	-	-	**91.8±2.6**	93.5±1.4	87.3±1.0	90.2±0.8
GDRO (Sagawa et al., 2020)	✓	✓	78.5±4.5	90.6±0.1	89.9±0.6	92.0±0.6	**88.9±1.3**	**93.9±0.1**
DISC (Wu et al., 2023)	✓	✓	-	-	88.7±0.4	93.8±0.7	-	-
GEORGE (Sohoni et al., 2020)	×	✓	76.4±2.3	89.5±0.3	76.2±2.0	**95.7±0.5**	54.9±1.9	**94.6±0.2**
LfF (Nam et al., 2020)	×	✓	0.0±0.0	25.0±0.0	78.0	91.2	77.2	85.1
CIM (Taghanaki et al., 2021)	×	✓	0.0±0.0	36.8±1.3	77.2	95.6	83.6	90.6
JTT (Liu et al., 2021)	×	✓	74.5±2.4	90.2±0.8	83.8±1.2	89.3±0.7	81.5±1.7	88.1±0.3
CnC (Zhang et al., 2022)	×	✓	77.4±3.0	90.9±0.6	88.5±0.3	90.9±0.1	88.8±0.9	89.9±0.5
SPARE (Yang et al., 2024)	×	✓	**83.0±1.7**	91.8±0.7	89.8±0.6	94.2±1.6	**90.3±0.3**	91.1±0.1
AFR (Qiu et al., 2023)	×	✓	-	-	90.4±1.1	94.2±1.2	82.0±0.5	91.3±0.3
CoBalT$_{hg}$ (ours)	×	✓	79.0±4.3	**96.6±1.8**	**90.6±0.7**	93.7±0.6	88.0±2.5	92.3±0.7
ERM	×	×	0.0±0.0	20.1±0.2	62.6±0.3	**97.3±1.0**	47.7±2.1	**94.9±0.3**
MaskTune (Asgari et al., 2022)	×	×	-	-	86.4±1.9	93.0±0.7	78.0±1.2	91.3±0.1
ULA (Tsirigotis et al., 2023)	×	×	**75.1±0.8**	-	86.1±1.5	91.5±0.7	86.5±3.7	93.9±0.2
XRM (Pezeshki et al., 2023)	×	×	70.5	-	86.1	90.6	**89.8**	91.8
CoBalT$_{ig}$ (ours)	×	×	73.5±2.1	96.0±1.6	89.0±1.6	92.5±1.7	89.2±1.2	92.3±0.6
CoBalT$_{avg}$ (ours)	×	×	74.5±2.0	**96.2±2.0**	**90.6±0.7**	93.8±0.8	81.1±2.7	92.8±0.9

nearly 3% on Waterbirds, while also demonstrating competitive performance on CelebA and CMNIST with an average accuracy similar to the other methods.

Even when selecting the model based on the average validation accuracy (CoBalT$_{avg}$), without attempting to infer groups, our model still demonstrates competitive results. Unlike other baselines that leverage human-annotated group-labeled training or validation sets for early stopping or hyper-parameter tuning (as detailed in Appendix C), our method makes group inferences for both training and validation data without relying on human labels.

Furthermore, our method exhibits similar performance to other methods employing group annotations. We provide visualizations of the feature attributions of ERM and our method in the Waterbirds dataset, as illustrated in Figure 4, demonstrating that our method relies less on spurious backgrounds compared to ERM.

4.2.2. Scenario 2 (Urban Cars)

When multiple spurious correlations are present in the datasets, existing methods show a Whack-a-mole behavior, where mitigating one spurious correlation amplify the other in minority groups. *CoBalT* is good at identifying abstractions from data, which help to identify multiple spurious attributes and thus improve spurious correlation.

From Table 2, we can see that when there are two spurious correlations such as backgrounds and co-occurring objects, our method overcomes these correlations and achieves SOTA performance outperforming *SPARE* by 3%. We also see that methods like *EIIL*, *LfF* and *JTT* struggle to perform better in the presence of multiple spurious correlations.

Table 2. Results from the Urban Cars dataset using ResNet-50, when there are multiple spurious correlations (BG+CoObj) exist.

Method	Worst Group	Average
ERM	28.4	**97.6**
EIIL	50.6	95.5
GEORGE	35.2	97.9
LfF	34.0	97.2
JTT	55.8	95.9
SPARE	76.9±1.8	96.6±0.5
GDRO	75.2	91.6
CoBalT$_{ig}$ (ours)	**80.0±2.8**	96.3±0.6
CoBalT$_{avg}$ (ours)	76.8±6.5	97.3±0.7

4.2.3. Scenario 3 (ImageNet-9 Background)

In the more realistic setting of the ImageNet-9 background challenge dataset, we assess the attribute generalization capability of our method. Training our model exclusively on the original ImageNet-9 trainset, without accessing the 'mask-rand' subset where background images are randomly swapped, we select the model based on inferred worst group performance on the original validation set.

As illustrated in Table 3, our method (CoBalT$_{ig}$) outperforms MaskTune (Asgari et al., 2022) by 1.1% on Mixed-same, 1.5% on Mixed-rand, and 1.9% on Only-FG. Additionally, we observe improvements compared to other methods across all test sets. These results underscore the efficacy of our concept discovery method and the importance weight-based sampling strategy in learning task-relevant information and mitigating spurious correlations. Notably, our sampling technique for addressing imbalances within the training set remains effective even in scenarios where the imbalance is not readily apparent.

Many techniques used for Waterbirds and CelebA are inapplicable to this dataset due to the lack of inference groups. Our method, however, is more versatile and performs well across various scenarios.

Table 3. Results on Background Challenge (ImageNet-9). Top rows based on ResNet-50 (ImageNet-Pretrained), 4 slots and codebook size 8. The results of other methods are taken from Asgari et al. (2022).

Method	Original	Mixed-same	Mixed-rand	Only-FG
ERM	**97.9**	90.5	79.2	88.5
CIM (Taghanaki et al., 2021)	97.7	89.8	**81.1**	-
SIN (Sauer & Geiger, 2021)	89.2	73.1	63.7	-
INSIN (Sauer & Geiger, 2021)	94.7	85.9	78.5	-
INCGN (Sauer & Geiger, 2021)	94.2	83.4	80.1	-
MaskTune (Asgari et al., 2022)	95.6	91.1	78.6	88.1
CoBalT$_{ig}$ (ours)	**97.9**	**91.2**	80.1	90.0
CoBalT$_{avg}$ (ours)	**97.9**	**91.2**	80.3	**90.1**

4.2.4. Results without validation groups

In our previous evaluations, we selected the model by early stopping based on the worst group validation performance, with the groups being inferred on the validation data by our proposed method. To evaluate the impact of model selection, we now consider a scenario where we lack access to human-annotated validation groups for CelebA. In this case, other methods select the model based on average validation accuracy, as they typically rely on human-annotated validation groups.

From Table 4, we can see that the performance of different methods substantially degrades when group-labeled validation data is unavailable for early stopping. Many of the group inference methods perform even worse than ERM, with the notable exception of MaskTune. However, MaskTune still performs significantly worse than our methods CoBalT$_{ig}$ and CoBalT$_{avg}$. This underscores the critical importance of having access to group-labeled data for many baseline methods to effectively work.

In contrast, our method proves valuable by inferring groups in an unsupervised manner. when we perform early stopping based on average validation accuracy, akin to the baseline methods in this table, our method CoBalT$_{avg}$ significantly outperforms others, particularly on the worst group.

Table 4. Results from the CelebA dataset using ResNet-50 (when early stopping is not done using validation group labels for other methods). We do early stopping based on our inferred groups on the validation set without using validation group labels. The baseline results are taken from Asgari et al. (2022).

Method	Worst Group	Average
ERM	47.7±2.1	**94.9±0.3**
CVaR DRO (Levy et al., 2020)	36.1	82.5
DivDis (Lee et al., 2023)	55.0	90.8
LfF	24.4	85.1
JTT	40.6	88.0
MaskTune	78.0±1.2	91.3±0.1
CoBalT$_{ig}$ (ours)	**89.2±1.2**	92.3±0.6
CoBalT$_{avg}$ (ours)	81.1±2.7	92.8±0.9

4.2.5. Interpretation of concepts

Our proposed method decomposes images into high-level concepts in an unsupervised way and clusters the images based on those concepts. Through the slot-based decomposition model, objectness or high-level concepts emerge in complex real-world data sets, which can be viewed through the attention map of each slot as in Figure 3. For example, in Waterbirds, the region grouped by slots belongs to parts of the body of birds and background such as trees, water, etc. In the IN-9L dataset, the slot distinguishes humans, animals, grass, etc. For CelebA, the model learns to separate the nose, eyes, and hair on the human face.

These decomposed slot representations are matched with a set of vector-quantized codes from the learned dictionary. Each code in the dictionary represents high-level abstract concepts. This approach effectively makes each code as the

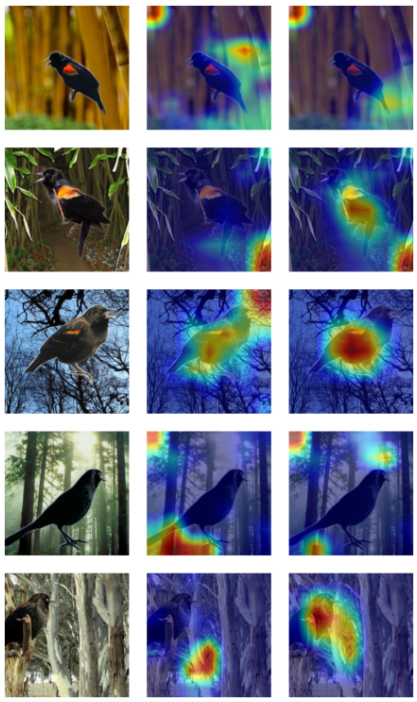

Figure 4. Gradcam heat-map on Waterbirds dataset (from left to right Input, ERM and CoBalT respectively in three columns). ERM models spuriously correlates to background information for classifying bird types whereas our methods reduce the spurious correlation by focusing on image regions that contain birds.

centroid of a specific cluster. By matching slot representations to the closest centroid, we can categorize an image into multiple distinct clusters. This allows us to identify and group images based on shared high-level concepts, despite the fact that they belong to different classes. Such an organization becomes particularly insightful when we observe images from varied classes clustering together. This clustering is based on the similarity of the underlying concepts these images represent. For example, images from different classes but with a common feature or concept might find themselves grouped in the same cluster (e.g. trees and water in Figure 1 respectively).

4.2.6. Limitations without Spatially Separable Concepts

Our method relies on the disentanglement of different concepts by clustering slots, which implicitly assumes spatial separability between different concepts. These concepts are assumed to be at the object level that might represent foreground/background spurious correlations. We hypothesize that because of this, on CMNIST and CelebA (Table 1), where concepts are less spatially separable, our method's benefit is a bit limited compared to other datasets like Waterbirds, Urban Cars, Bar (Nam et al., 2020) (Table 6), where concepts have clear spatial regions. Further investigation in this direction is needed. In principle, one could try to learn a disentangled representation of the high level objects and use the disentangled factors of variations as concept. Work like Singh et al. (2022) that further factorizes the representation of each semantic region can be used to mitigate this issue by allowing one spatial region to be represented as a collection of concepts.

4.2.7. Avoiding Representation Collapse

Our concept discovery method is based on self-supervised Siamese representation learning, utilizing two parallel encoders: the student produces the source slot encoding and the teacher produces the target encoding. One of the main issues with this kind of encoder-only learning framework is representation collapse (Hua et al., 2021). During training, our method can obtain a degenerate solution in which all representations of the slots fall into one cluster, while still minimizing the objective in Equation 5.

To avoid this degenerate case, we employ a similar set of ideas as DINO (Caron et al., 2021) to have asymmetric teacher and student branches: 1) using data augmentations of teacher and student views; 2) centering and sharpening of teacher slot distributions; 3) updating teacher weights by taking an exponential moving average of student. Typically, the teacher model's weights are updated after every gradient update step for most datasets. However, for the CMNIST datasets, data augmentation is not used. To maintain the asymmetry between the teacher and student models in the absence of data augmentation, the updates for the CMNIST datasets are performed less frequently, specifically after every 20 steps.

5. Conclusion

Drawing inspiration from object-centric representation learning based on slot attention, we proposed a framework for decomposing images into concepts in an unsupervised way. We demonstrated the effectiveness of these concept clusters in discerning between minority and majority group samples within the dataset, all without relying on human group annotations. Leveraging these concepts, we devised an importance sampling technique that prioritizes rare concepts for each class, culminating in the training of a robust model exhibiting consistent performance with existing baselines in mitigating worst group errors.

Our exploration in this paper has been confined to vision datasets; however, future investigations could extend to NLP or multi-modal datasets to further alleviate biased learning. Additionally, promising research avenues involve techniques targeting spurious concepts, such as concept-aware data augmentations, warranting further exploration.

Acknowledgements

We acknowledge the support of the Canada CIFAR AI Chair Program and IVADO. We thank Mila and Compute Canada for providing computational resources.

Impact Statement

This paper leverages the discovery of unsupervised concepts to train robust classifiers by reducing spurious correlations without explicit human annotations of spurious attributes. This research seeks to reduce biases in model training and improve decision-making processes across various applications, potentially benefiting areas such as healthcare, autonomous systems, and content moderation. We acknowledge the importance of considering ethical implications and societal impacts, especially in ensuring fairness and transparency in automated decisions. Future developments should continue to address these concerns, promoting equitable and responsible AI.

References

Alain, G., Lamb, A., Sankar, C., Courville, A., and Bengio, Y. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

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

Asgari, S., Khani, A., Khani, F., Gholami, A., Tran, L., Mahdavi-Amiri, A., and Hamarneh, G. Masktune: Mitigating spurious correlations by forcing to explore. In *Advances in Neural Information Processing Systems*, 2022.

Azulay, A. and Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019. URL http://jmlr.org/papers/v20/19-519.html.

Bahng, H., Chun, S., Yun, S., Choo, J., and Oh, S. J. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pp. 528–539. PMLR, 2020.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf.

Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. URL https://api.semanticscholar.org/CorpusID:233444273.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Cho, J. H., Mall, U., Bala, K., and Hariharan, B. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, 2021.

Engelcke, M., Jones, O. P., and Posner, I. Genesis-v2: Inferring unordered object representations without iterative refinement. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:233307216.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation

learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Hua, T., Wang, W., Xue, Z., Wang, Y., Ren, S., and Zhao, H. On feature decorrelation in self-supervised learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9578–9588, 2021. URL https://api.semanticscholar.org/CorpusID:233481690.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In Schölkopf, B., Uhler, C., and Zhang, K. (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 336–351. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/idrissi22a.html.

Kim, E., Lee, J., and Choo, J. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14992–15001, 2021.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, Y., Yao, H., and Finn, C. Diversify and disambiguate: Out-of-distribution robustness via disagreement. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=RVTOp3MwT3n.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

Li, Z., Evtimov, I., Gordo, A., Hazirbas, C., Hassner, T., Ferrer, C. C., Xu, C., and Ibrahim, M. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20071–20082, 2023.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2014.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.

Moayeri, M., Wang, W., Singla, S., and Feizi, S. Spuriosity rankings: Sorting data to measure and mitigate biases. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=jSuhnO9QJv.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684, 2020.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_F9xpOrqyX9.

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

Pezeshki, M., Bouchacourt, D., Ibrahim, M., Ballas, N., Vincent, P., and Lopez-Paz, D. Discovering environments with xrm. *arXiv:2309.16748 [cs.LG]*, 2023.

Qiu, S., Potapczynski, A., Izmailov, P., and Wilson, A. G. Simple and fast group robustness by automatic feature reweighting. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28448–28467. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/qiu23c.html.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Roy, A., Vaswani, A., Parmar, N., and Neelakantan, A. Towards a better understanding of vector quantized autoencoders. 2018.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Sauer, A. and Geiger, A. Counterfactual generative networks. *ArXiv*, abs/2101.06046, 2021. URL https://api.semanticscholar.org/CorpusID:231627872.

Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C.-J., He, T., Zhang, Z., Schölkopf, B., Brox, T., and Locatello, F. Bridging the gap to real-world object-centric learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=b9tUk-f_aG.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33: 9573–9585, 2020.

Singh, G., Kim, Y., and Ahn, S. Neural systematic binder. In *International Conference on Learning Representations*, 2022. URL https://api.semanticscholar.org/CorpusID:255749563.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Taghanaki, S. A., Choi, K., Khasahmadi, A. H., and Goyal, A. Robust representation learning via perceptual similarity metrics. In *International Conference on Machine Learning*, pp. 10043–10053. PMLR, 2021.

Tsirigotis, C., Monteiro, J., Rodriguez, P., Vazquez, D., and Courville, A. Group robust classification without any group information. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=2OcNWFHFpk.

van den Oord, A., Vinyals, O., and kavukcuoglu, k. Neural discrete representation learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

Wen, X., Zhao, B., Zheng, A., Zhang, X., and Qi, X. Self-supervised visual representation learning with semantic grouping. *Advances in Neural Information Processing Systems*, 35:16423–16438, 2022.

Wu, S., Yuksekgonul, M., Zhang, L., and Zou, J. Discover and cure: Concept-aware mitigation of spurious correlation. In *International Conference on Machine Learning*, pp. 37765–37786. PMLR, 2023.

Xiao, K. Y., Engstrom, L., Ilyas, A., and Madry, A. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gl3D-xY7wLq.

Yang, Y., Gan, E., Dziugaite, G. K., and Mirzasoleiman, B. Identifying spurious biases early in training through the lens of simplicity bias. In *International Conference on Artificial Intelligence and Statistics*, pp. 2953–2961. PMLR, 2024.

Yong, L., Zhu, S., Tan, L., and Cui, P. ZIN: When and how to learn invariance without environment partition? In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=pUPFRSxfACD.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Zhang, M., Sohoni, N. S., Zhang, H., Finn, C., and Re, C. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations, 2022. URL https://openreview.net/forum?id=cVak2hs06z.

Zhang, Y., Zhang, D. W., Lacoste-Julien, S., Burghouts, G. J., and Snoek, C. G. M. Unlocking slot attention by changing optimal transport costs. In *International Conference on machine Learning*, 2023.

A. Detailed architecture

Encoders and Projectors. We employ a *student encoder* f_s, *projector* g_s for a branch we call the *student branch* which is updated using stochastic gradient descent (SGD). For simplicity, we represent the parameter weights of both the encoder and the projector of the student branch by θ_s. Similarly, the other branch named the *teacher branch*, parameterised by θ_t, has the same set of architectural components, respectively, *teacher encoder* f_t, *projector* g_t, but with different sets of weights, which are updated by the exponential moving average of the student parameters as:

For an input image, two randomly augmented views $x_s, x_t \in R^{c \times h \times w}$ are created, where c, h, and w are input image channel, height and width respectively; the goal is to extract slot representations from one view of the teacher branch and apply consistency to another view of the student branch. Firstly, the augmented views are encoded and projected by the student and teacher encoder and projectors as $y_s = g_s(f_s(x_s)) \in R^{P \times d}$ and $y_t = g_t(f_t(x_t)) \in R^{P \times d}$, where are P spatial patch representations of dimension d.

Slot Module. We can learn abstract concept representations by grouping semantically similar patch representation vectors. To do that, we employ randomly initialized vectors for the student branch, which we call *student slots* $S_s \in R^{N \times d}$, where N is the number of slots with dimension d. These slot vectors are then used to perform an attention-weighted pooling of patch vectors. To encourage competition among slots, where each slot attends to distinctive and semantically similar patch vectors, we utilize slot attention (Locatello et al., 2020) where attention is calculated over the 'slot' axis as below:

We define any normalized vector as $\bar{z} = z/||z||$ *and any column normalized matrix as* $\bar{Z} = [\bar{z}_1, ..., \bar{z}_N]$

$$A_s = \underset{N}{\text{soft}}\max(\bar{S}_s.\bar{y_s}^T/\tau_s) \in R^{N \times P} \tag{7}$$

where τ_s is the temperature for the student.

Then we can calculate **concept representations** using slots by pooling the patch representations based on the attention maps as follows:

$$z_s = A_s.y_s, \quad z_t = A_t.y_t \in R^{N \times d} \tag{8}$$

Teacher We do the same for the teacher branch, *teacher slots* by initializing them from the student slots weights $S_t \in R^{N \times d}$ as: The weights of S_s are updated by SGD, but the weights of S_t are updated by the exponential moving average of S_s as follows:

$$\theta_t = \alpha_t \cdot \theta_s + (1 - \alpha_t) \cdot \theta_t \tag{9}$$

$$S_t = (1 - \alpha_t) \cdot S_s + \alpha_t \cdot S_t \tag{10}$$

$$A_t = \underset{N}{\text{soft}}\max(\bar{S}_t.\bar{y_t}^T/\tau_t) \in R^{N \times P} \tag{11}$$

where τ_t is the temperature for the teacher.

Loss Since we utilize a two-branch teacher-student architecture for better inductive bias to facilitate higher-level abstraction, we focus on attention distillation from teacher to student. The student model learns to mimic the attention patterns of the teacher, effectively capturing the representation of the essential abstract concept from the data without explicitly reconstructing the input.

We utilize the attention distillation loss introduced in Wen et al. (2022):

$$\boldsymbol{\mathcal{L}_{dis}} = -\sum_{N}\sum_{P} M \circ A_t \log A_s \tag{12}$$

where M is a mask that prevents distillation of non-overlapping patches from the teacher to the student views. Since the student and teacher branches observe two randomly cropped views, there may exist non-overlapping patches to which we do not want to apply distillation.

To avoid redundant slots and facilitate the learning of different information, we use a contrastive loss between the slot representations introduced in Wen et al. (2022). This ensures that similar concepts have closer representations in the embedding space, while dissimilar ones are further apart.

$$\boldsymbol{\mathcal{L}_{con}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{I(i)\exp\left(p(\bar{z}_s^i).\bar{z}_t^i/\tau_c\right)}{\sum_{\hat{i}=1}^{N} I(\hat{i})\exp\left(p(\bar{z}_s^{\hat{i}}), \bar{z}_t^{\hat{i}}/\tau_c\right)} \tag{13}$$

where p is predictor network as in (Caron et al., 2021) and I is an indicator function that finds common slots between the views of the teacher and the student after masking out the slots that fail to attend to any patch as below:

$$\mathbf{I}(i) = \begin{cases} 1 & \text{if } (\mathbf{m}_s)_i = (\mathbf{m}_t)_i \\ 0 & \text{otherwise} \end{cases}$$

$$(\mathbf{m}_s)_i = \sum_{j=1}^{P} \mathbb{1}\{i = \text{argmax } A_s^{:j}\} \geq 1$$

$$(\mathbf{m}_t)_i = \sum_{j=1}^{P} \mathbb{1}\{i = \text{argmax } A_t^{:j}\} \geq 1$$

B. Implementation Details

Architecture

For the training of the concept learning model, Imagenet pre-trained ResNet-50 has been used for the student encoder f_s and the teacher encoder f_t. The student p_s and teacher p_t projector networks are similar to(Caron et al., 2021) with a hidden dimension of 1024 and an output dimension of 32. We also use 32 as the slot dimension and the concept vector dimension for all datasets except CMNIST. For CMNIST, we use slot and hidden dimensions of 16 and 32, respectively.

B.1. Training Details

We train the concept discovery and classification model using ResNet50 (He et al., 2016) pre-trained on ImageNet from the Pytorch library (Paszke et al., 2019) as a backbone for all data sets except CMNIST where LeNet-5 (LeCun et al., 1998) is used. All experiments were performed with NVIDIA A100 and V100 GPUs.

Data Augmentation

For the training of the concept learning model, we follow the data augmentation scheme proposed in Wen et al. (2022).

Hyperparameters

For both the student and teacher networks, the temperature values τ_s and τ_t are 0.1 and 0.07, similar to Caron et al. (2021). For the exponential moving average (EMA) update coefficients α_c and α_t, we use 0.9 and 0.99 respectively. To obtain the sampling factor (λ) in Algorithm 1, we use the average silhouette score as in Yang et al. (2024). For all datasets, when the average silhouette score ≤ 0.8, we set $\lambda = 2$, otherwise 1. For Waterbirds and Urban Cars, it is 2, and for other data sets, it is set to 1.

For training the concept discovery model, we use Adam (Kingma & Ba, 2015) as an optimizer with a learning rate of $2e^{-4}$ and a weight decay of $5e^{-4}$ for 50 epochs with a batch size of 128. The same configuration is used for all data sets, except CMNIST and CelebA, which are trained for 20 epochs. For CMNIST, the batch size is 32.

We train classification models using SGD with 0.9 momentum for all datasets. The learning rate is $1e^{-4}$, except for CMNIST ($1e^{-3}$). A weight decay of 0.1 is applied to Waterbirds, CelebA, and Urban Cars. Training epochs: Waterbirds (300), CelebA (60), IN-9L (100), Urban Cars (300), and CMNIST (20). The batch size is 128 for all datasets, except CMNIST (32).

C. Additional Results

Ablation Studies

To perform an ablation study on the **Waterbirds** dataset, we used different numbers of slots $\{2, 4, 6, 8\}$ and the size of the codebook $\{2, 4, 8, 12, 16\}$. From Figure 6, we can see the segmentation mask with a varying number of slots. From the figure, it is clear that on average $3-4$ and $2-6$ slots are activated per image when we initialize with 4 and 6 slots, respectively. We can identify the activated slots for each image based on Equation A. If we increase the number of slots, we can see fine-grained segmentation. Figure 7 and Figure 8 shows similar ablation studies of segmented images while varying the # slots in the **CelebA** and **IN-9L** data sets, respectively.

From Figure 5, we can see the impact on the performance of the worst group accuracy when varying the number of slots and the size of the codebook. It is evident that over-segmentation (with more slots) degrades the worst group accuracy. We hypothesize that we can have the right balance between the worst group and the average accuracy with a suitable number of slots, which encourages us to learn abstract semantic concepts. In all of our experiments, we use 4 as the # slot for all data sets.

Table 5 shows the worst group and average accuracy on **CelebA** dataset while varying the size of the codebook and fixing the # slots to 4.

Codebook Size	Worst group Acc	Avg Acc
4	74.4	93.5
8	88.3	92.6
12	86.7	89.7
16	88.3	90.4

Table 5. Varying Codebook size on CelebA (# slots fixed to 4)

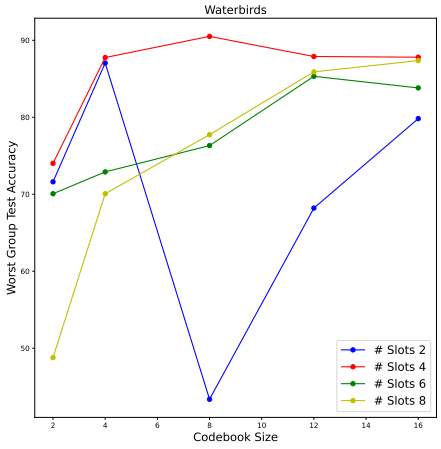

Figure 5. Ablation on Worst group Test Accuracy on Waterbirds dataset with varying slots and codebook size

Results on Bar Dataset with Spatially Separable Concepts

We further evaluated our method on the Biased Action Recognition (BAR) dataset (Nam et al., 2020), which includes six action classes biased to distinct places. In this dataset, the concepts have clear spatial regions and we see (Table 6) the efficacy of our method, where it outperforms other baselines.

Table 6. Result on BAR (Nam et al., 2020) dataset

ERM	ReBias (Bahng et al., 2020)	LfF	BiaSwap (Kim et al., 2021)	CoBalT$_{ig}$
51.9±5.9	59.7±1.5	63.0±2.8	52.4	**67.0±0.9**

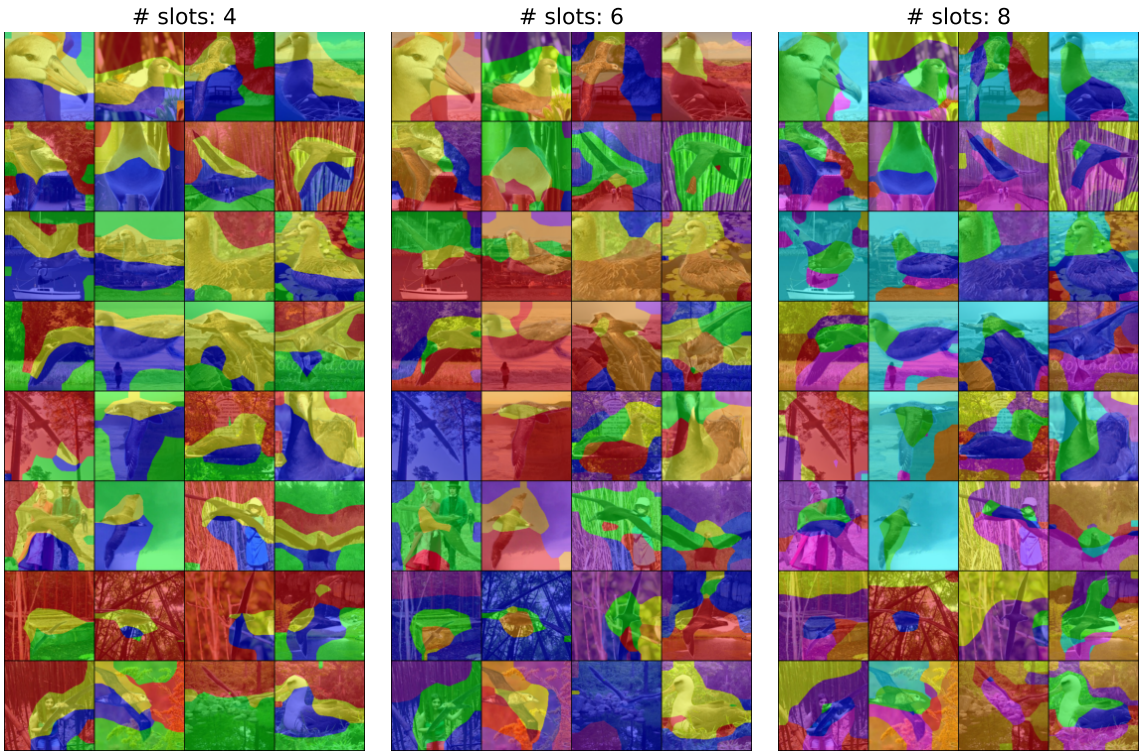

Figure 6. Ablation of segmentation mask with varying slots on Waterbirds

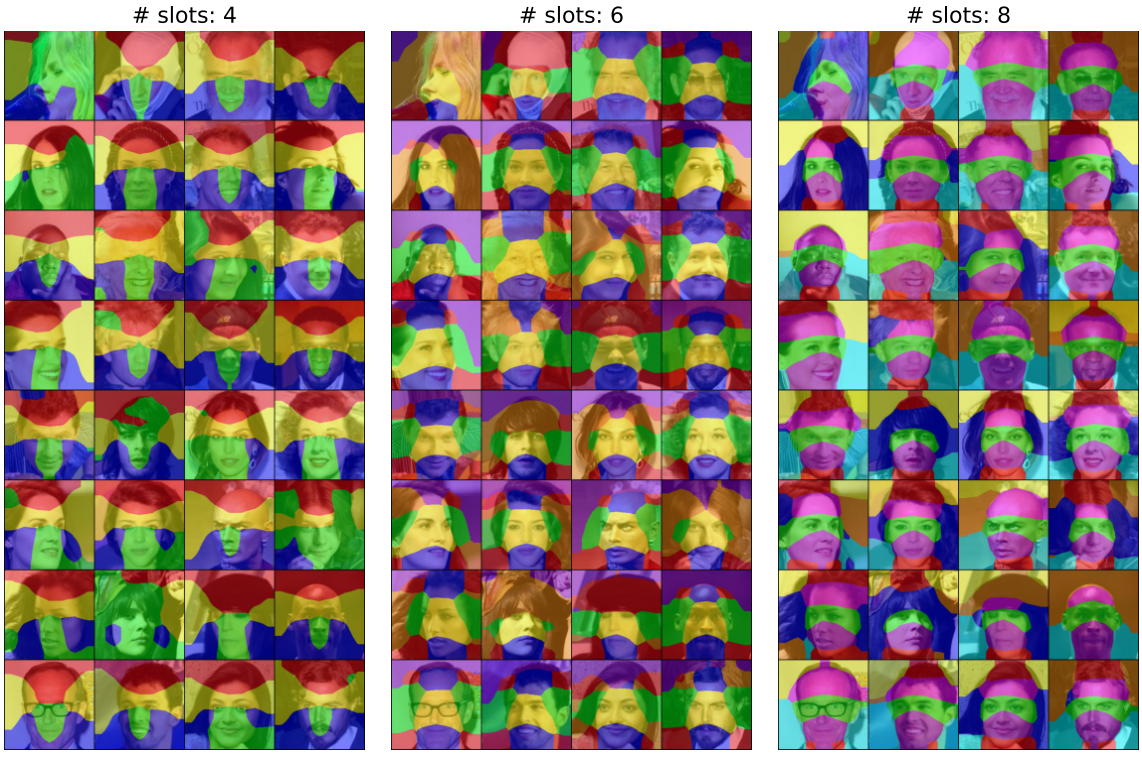

Figure 7. Ablation of segmentation mask with varying slots on CelebA

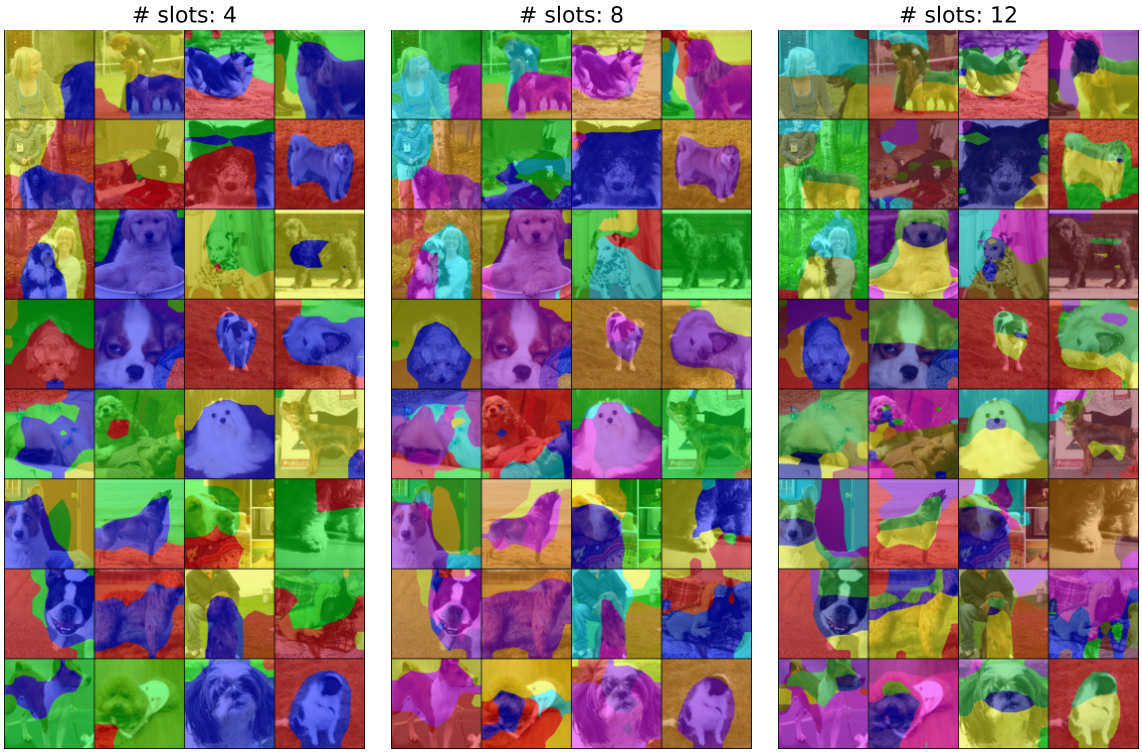

Figure 8. Ablation of segmentation mask with varying slots on IN-9L