Group-wise Contrastive Bottleneck for Weakly-Supervised Visual Representation Learning

Boon Peng Yap, Beng Koon Ng
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
{boonpeng001, ebkng}@ntu.edu.sg

Abstract

Coarse or weak labels can serve as a cost-effective solution to the problem of visual representation learning. When fine-grained labels are unavailable, weak labels can provide some form of supervisory signals to guide the representation learning process. Some examples of weak labels include image captions, visual attributes and coarse-grained object categories. In this work, we consider the semantic grouping relationship that exists within certain types of weak labels and propose a group-wise contrastive bottleneck module to leverage this relationship. The semantic group may contain labels that are related to a general concept, such as the colour or shape of objects. Using the group-wise bottleneck module, we disentangle the global image features into multiple group features and apply contrastive learning in a group-wise manner to maximize the similarity of positive pairs within each semantic group. The positive pairs are defined based on the similarity of the labels captured by each group. To learn a more robust representation, we introduce a reconstruction objective where an image feature is reconstructed back from the disentangled features, and this reconstruction is encouraged to be consistent with the feature obtained from a different augmented view of the same image. We empirically verify the efficacy of the proposed method on several datasets in the context of visual attribute learning, fair representation learning and hierarchical label learning. The experimental results indicate that our proposed method outperforms prior weakly-supervised methods and is flexible in adapting to different representation learning settings.

1. Introduction

Self-supervised learning has received a lot of research interest from the computer vision [3, 4, 8, 10, 11] and natural language processing [1, 7] community, largely due to its ability to extract generic representations from potentially unlimited annotation-free data. Recent advances in self-supervised learning have focused on two main approaches: contrastive learning and masked reconstruction. In contrastive learning, positive pairs are defined or constructed, and the learning objective is to maximize their similairity in a shared representation space. On the other hand, masked reconstruction applies random masking to the inputs and learns to reconstruct the original inputs using a decoder. After pretraining on large-scale training data, pretrained self-supervised models can be transferred to a wide range of downstream tasks and in some cases achieve performance similar to or even surpass that of fully supervised models.

In this work, we consider the problem of visual representation learning on small to medium-scale training data, specifically by exploring how to incorporate additional information provided by weak/auxiliary annotations. In the real world, images are often annotated with auxiliary labels that are related to the visual content. For example, images posted on social media are usually accompanied by hashtags consisting of keyword phrases to facilitate engagement with other users. Such labels exist abundantly and are continuously growing as more content is being uploaded every day, but they may be too noisy and require manual cleaning before they can be used as pretraining data [20]. Another form of weak labels are the visual attributes such as the color and shape attributes. They required some manual labeling but it is often much easier than identifying the exact object class. For example, in the task of classifying bird species, visual attributes such as the beak color can be inferred without much thought compared to the species type, which may require more specialized knowledge. While imprecise, weak labels can still provide valuable training signals for learning better visual representations, especially in scenarios where fully annotated data is limited.

Prior work in weakly-supervised representation learning has proposed to construct positive pairs for contrastive learning using weak labels. One straightforward way is to define positive pairs as instances having similar labels, *e.g.*, using the spatial and time information from camera meta-

data [23]. However, this approach may not be applicable to cases where the labels are sparse and have a large dimension. Due to the sparsity, only a few instances will have matching labels, resulting in fewer positive pairs. To mitigate this issue, clustering InfoNCE [31] ranks each class label by its entropy and selects the top-k classes to cluster the data. Data points which share the same labels in the selected top-k classes are assigned to the same cluster, and the positive pairs are then sampled from these clusters. Instead of discarding potentially useful information, another approach utilizes a similarity kernel (*e.g.* cosine similarity) to compute the similarity score between two sets of labels [32]. The similarity scores are used to weight the loss for any given pair of contrasting samples.

Although effective, prior approaches does not explicitly consider the semantic relationship between the class labels. In a hierarchical relationship, class labels are organized in a tree-like structure, where the top-level class subsumes the classes below it. This relationship has previously been explored in previous work [40], which adapted the contrastive loss to a hierarchical setting. In this work, we investigate another type of semantic relationship where the labels are related to some attribute groups. For example, the labels "*red*", "*green*", and "*blue*" belong to the color attribute group, while the labels "*square*" and "*circle*" belong to the shape attribute group. An image may be annotated with labels from multiple groups, and two images may share similar labels in some of the groups (*e.g.* a red square and a red circle share the same label in the color attribute group but not the shape attribute group). To model this relationship, we propose a group-wise bottleneck learning mechanism in which the global representation of an image is split into multiple group-specific representations. Each group-specific representation is then projected into a sub-space corresponding to an attribute group. As shown in Fig. 1, within each sub-space, positive pairs are defined for contrastive learning based on the shared labels that belong to the attribute group associated with that sub-space. This ensures sufficient amount of positive pairs within each group as the visual attributes are modeled on a per-group basis. In addition, through a reverse mapping operation similar to how an autoencoder [18] works, a global representation is reconstructed from the group-wise representations. The reconstructed representation is compared against another representation extracted from an augmented view of the same image, and their discrepancy is minimized through a mean square error loss. This produces a more robust representation and encourages learning of residual features (i.e., unobserved attributes). We refer to the proposed approach as CLRC (**C**ontrastive **L**earning with **R**e**C**onstruction loss). The efficacy of CLRC is validated on multiple public datasets annotated with different visual attributes. In addition, we also demonstrate how

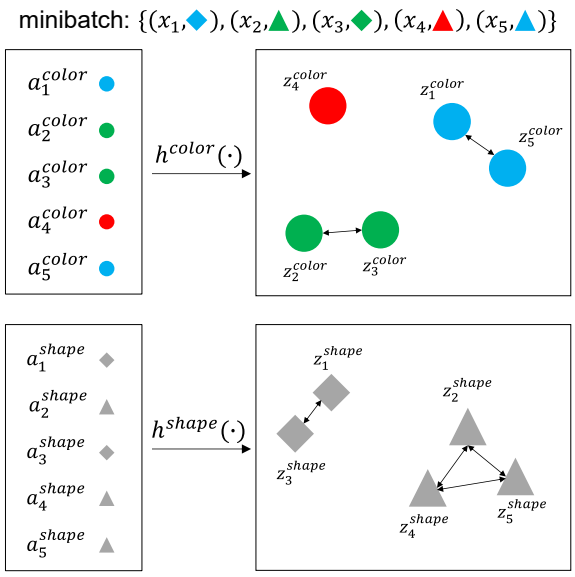

Figure 1. Illustration of the group-wise contrastive learning approach on a toy dataset. Each sample x in the minibatch is annotated with a color and shape attribute. The features a are mapped by group-specific projector h into projection vectors z. The double ended arrows indicate positive pairings between two samples.

CLRC can be adapted for hierarchical labels and fair representation learning. The experimental results show that our proposed approach outperforms prior approaches and is useful for mitigating the issue of subgroup bias by learning a fairer representation. The codes are made publicly available at https://github.com/BPYap/CLRC.

The main contributions of this work are threefold:

1. We introduce a bottleneck module with group-wise contrastive learning objective for weakly-supervised representation learning. The proposed module separately models the features based on label similarity in different semantic groups.
2. We propose to integrate group-wise contrastive learning with a feature reconstruction objective to improve the robustness of learnt representations.
3. We conduct extensive experiments to demonstrate the effectiveness of our proposed CLRC method in learning representations from visual attributes, hierarchical labels and sensitive attributes.

2. Related Work

Contrastive Learning. The goal of contrastive learning is to minimize the discrepancy between two semantically similar images while maximizing the discrepancy for semantically different images. The notion of semantic similarity depends on the availability and granularity of labels.

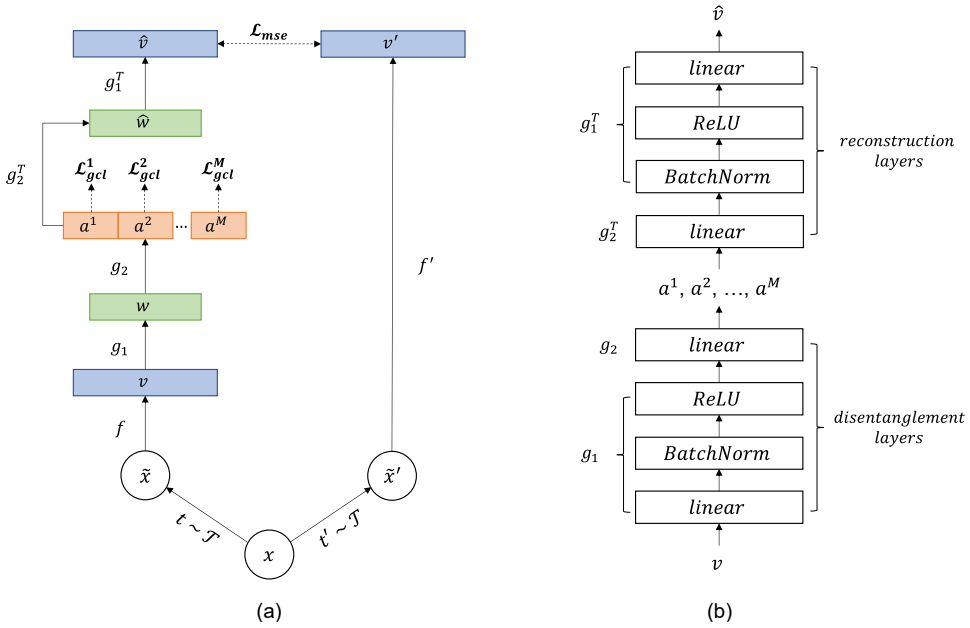

Figure 2. Overview of the proposed approach. (a): Two augmented views ($\tilde{x}, \tilde{x}'$) are generated from an input image x using transformations sampled from $\mathcal{T}$. The global representation v obtained from an encoder f is disentangled into group-specific representation a through the bottleneck module g. Using a set of small linear projectors (ommited in this diagram for brevity), a is separately projected into different sub-spaces where the group-wise contrastive objective $\mathcal{L}_{gcl}$ is applied. The reconstructed global representation $\hat{v}$ is encouraged to be similar (via mean square error $\mathcal{L}_{mse}$) to the representation v' obtained from the momentum encoder f'. (b): Details of the bottleneck module. The reconstruction layers share the same weights as the disentanglement layers through the transpose of the weights of g.

In self-supervised setting where manually annotated labels are not available, positive pairs of semantically similar images are typically generated from the same image through image augmentations [4, 8, 11]. In scenarios where some form of supervisory signal is available, the positive pairs can be defined as images sharing the same labels. For example, in supervised contrastive learning [16], in addition to the augmented images, different images with the same object category are also treated as positive pairs. Other forms of supervisory signal include visual attributes [31], hierarchical information [40], and camera metadata [23, 25]. In this work, our method utilizes the visual attributes and their semantic grouping information for weakly-supervised representation learning.

Weakly-Supervised Learning. To reduce the cost of acquiring precise annotations, weakly-supervised learning leverages weak labels that are imprecise or incomplete but are often significantly easier to acquire. It has been extensively studied in the literature of weakly-supervised semantic segmentation where weaker annotations such as bounding boxes [5, 26, 37] and image-level labels [14, 17, 36] are used in place of the costly pixel-wise segmentation masks. Beyond segmentation tasks, attribute labels [31, 32] and hash tags [20] have also been explored as a viable option for weak labels to improve performance on fine-grained classification tasks. In contrast to prior work which does not consider the semantic grouping of labels, our method explicitly models this relationship by disentangling the global representation into different semantic sub-spaces.

Feature Disentanglement In computer vision tasks such as image generation and image retrieval, learning representations that can be disentangled into interpretable components is highly desirable. This allows for better control in generating or retrieving images that meet certain criteria, *e.g.* changing the hair color of a person in a generated image by manipulating the part of the representation associated with hair color. A common approach to disentanglement learning is to train a separate classifier for each interpretable component through the cross-entropy loss [13, 39, 41]. Inspired by the idea of feature disentanglement, this work investigates whether the semantic grouping of visual attributes is useful for weakly-supervised representation learning and studies its synergy with the contrastive learning objective.

3. Methodology

An overview of the proposed method (CLRC) is presented in Fig. 2. CLRC involves disentangling the global

representation of an image into different semantic groups via a group-wise bottleneck module. Within each semantic group, contrastive learning objective is used to maximize the similarity of the projection vectors based on the shared labels in that group (Fig. 1). Then, the global representation is reconstructed from the disentangled representations and its consistency is maximized to match the representation extracted from a momentum encoder. The details of each step are described in the following sections.

3.1. Group-wise Contrastive Learning

This work considers datasets that are annotated with binary labels where each label indicates the presence of a visual attribute such as "*blue feather vs. not blue feather*" or "*has stripe vs. no stripe*". Naturally, some binary attributes could be grouped into a broader semantic category. For example, the attributes "*blue feather vs. not blue feather*" and "*red feather vs. not red feather*" belong to a semantic group related to the feather color. Based on this observation, a group-wise bottleneck module is introduced to exploit the semantic grouping information of binary attributes. As shown in Fig. 2(b), the bottleneck module consists of disentanglement and reconstruction layers. Given the global representation, v, of an image, the disentanglement layers produce a set of group-specific representations, $a^1, a^2, ..., a^M$, where M is the number of semantic groups. Concretely, the disentanglement layers include two linear layers, in which the first linear layer is followed by a Batch Normalization layer [15] and a ReLU activation function (all three layers will collectively be referred to as g_1). g_1 maps v, which has a feature dimension of d, into an intermediate representation w with a dimension of d_1 ($d_1 \leq d$), while the second linear layer, g_2, maps w into the group representation a with a dimension of $d_2 * M$ ($d_2 < d_1$). The group-specific features are then obtained by splitting a into M equal-size vectors.

To learn disentangled features, group-wise contrastive learning is proposed to separately maximize the similarities of positive pairs in different sub-spaces. Within each group, a light-weight linear projector is used to obtain the ℓ_2 normalized projection vectors, z, and the contrastive learning objective [4, 33] is optimized:

$$\mathcal{L}_{gcl}(z_i, z_j) = -\log \frac{\exp\left(z_i^T z_j/\tau\right)}{\sum_{k=1}^{N} \mathbb{1}(i \neq k) \exp\left(z_i^T z_k/\tau\right)}, \quad (1)$$

where z_i and z_j is a positive pair, defined in a group-wise manner as two instances that share at least one common attribute in the group, N is the batch size, and τ is a temperature parameter. The group-specific contrastive losses are then aggregated across all groups:

$$\mathcal{L}_{cl} = \frac{1}{\sum_{m=1}^{M} \lambda_m} \sum_{m=1}^{M} \frac{\lambda_m}{|B|} \sum_{(z_i^m, z_j^m) \in B} \mathcal{L}_{gcl}(z_i^m, z_j^m), \quad (2)$$

where B is the minibatch and λ_m is a weighting factor for the group-specific contrastive loss. An entropy-based weighting scheme is proposed to give more weightage to the more informative groups:

$$\lambda_m = -\sum_{\alpha \in A_m} p_\alpha \log p_\alpha, \quad (3)$$

where A_m is the set of binary attributes in the m-th group and p_α is the ratio of samples having attribute α in the training dataset.

3.2. Feature Reconstruction

To increase the robustness of learnt representations, the disentangled group representations are projected back to the global representation space through the reconstruction layers. The reconstruction layers adopt the tied weights design commonly used in autoencoders [34], where their weights are the transpose of the weights of the disentanglement layers. The reconstructed feature, $\hat{v}$ is aligned with the target feature v' extracted from a momentum encoder [11, 29] using a differently augmented input image. The weights of the momentum encoder f' are the exponential moving average version of the image encoder weights f, and no gradient is propagated to the momentum encoder during the computation of v'. The difference between $\hat{v}$ and v' is minimized via the mean square error function:

$$\mathcal{L}_{mse}(\hat{v}, v') = \frac{1}{d} \sum_{i=1}^{d} \left(\sigma(\hat{v}_i) - \sigma(v'_i)\right)^2, \quad (4)$$

where σ is a sigmoid function used to constrain the feature values within the range of $[0, 1]$, and i is an index into the feature vector. We posit that the reconstruction loss also encourages preservation of unannotated attributes, as the representations of important but unobserved attributes will be implicitly distributed over the group representations.

3.3. Overall Loss Function

During training, two augmented views are generated from each input image, and each view is involved in minimizing both $\mathcal{L}_{cl}$ and $\mathcal{L}_{mse}$. The overall loss function is a symmetrized version of the joint loss:

$$\mathcal{L}_{clrc} = \frac{\mathcal{L}_{cl}^1 + \mathcal{L}_{mse}(\hat{v}_1, v'_2) + \mathcal{L}_{cl}^2 + \mathcal{L}_{mse}(\hat{v}_2, v'_1)}{2}, \quad (5)$$

where $\mathcal{L}_{cl}^1$ (resp. $\mathcal{L}_{cl}^2$) denotes the application of $\mathcal{L}_{cl}$ on the minibatch of the first (resp. second) augmented views, $\hat{v}_1$ (resp. $\hat{v}_2$) denotes the reconstructed representation from the first (resp. second) augmented view, and v'_1 (resp. v'_2) denotes the target representation computed from the first (resp. second) augmented view.

Dataset	Type	#train-val	#test	#attribute	#group	#class
UT Zappos 50k [38]	visual attributes	35,017	15,008	126	6	21
WIDER Attribute [19]	visual attributes	6,871	6,918	14	14	30
CUB-200-2011 [35]	visual attributes	5,994	5,794	312	28	200
Fitzpatrick17k [9]	sensitive attributes	14,410	1,602	18	6	2
ImageNet-100 [6]	hierarchical labels	128,743	5,000	13	2	100

Table 1. Details of the benchmark datasets. #class is the number of object classes in the downstream tasks.

4. Experiments

Three sets of experiments are conducted to evaluate the efficacy of CLRC in handling different types of weak annotations, including grouped visual attributes, sensitive attributes and hierarchical labels. The details of the benchmark datasets are summarized in Tab. 1. Unless otherwise specified, the data preprocessing steps and image transformations follow the convention of previous works. The details are also included in the supplementary material.

4.1. Baselines

The proposed method is validated against a total of eight baseline methods, including three representative self-supervised methods (SimCLR [4], BYOL [8], SwAV [2]) to establish the lower bound performance, a supervised contrastive method with direct access to the downstream labels (SupCon [16]), and four baselines for weakly supervised learning. The first weakly supervised learning baseline is the cross-entropy supervision, which is trained to predict the concatenation of attribute vectors via the binary cross-entropy objective. The second baseline is based on the Contrastive Multiview Coding (CMC) [30] method, which involves maximizing the similarity between image representation and the representation of the attribute vectors obtained from a multilayer perceptron (MLP). The third method, Cl-InfoNCE [31], constructs positive pairs for contrastive learning by forming clusters using subsets of common attributes. The fourth method, CCL-K [32] uses a cosine similarity kernel on pairs of attribute vectors to compute the weights for the contrastive learning loss. For a fair comparison, all baselines are implemented under the same codebase using the PyTorch framework [24]. Our implementation produced different numbers than those reported by Cl-InfoNCE, possibly due to the differences in dataset-specific hyperparameters when training the linear classifier layer, such as the number of warmup and training epochs. As they were not explicitly stated in Cl-InfoNCE, we standardize the warmup and training epochs to 33 and 100, respectively, for the benchmarks in UT Zappos 50k, WIDER Attribute and CUB-200-2011. The full hyperparameter settings are provided in the supplementary material.

4.2. Learning from Grouped Visual Attributes

Datasets. Three visual attributes datasets are considered: 1) UT Zappos 50k [38], 2) WIDER Attribute [19], and 3) CUB-200-2011 [35]. For UT Zappos 50k and CUB-200-2011, the semantic groupings of the attributes are provided by the datasets. For WIDER Attribute, each binary attribute is treated as an individual semantic group.

Experiment Setup. The training and evaluation setup follow the protocols previously established in Cl-InfoNCE [31]. Hyperparameters and the encoder backbone are fixed for all experiments following the same settings in Cl-InfoNCE. For UT Zappos 50k, the encoder backbone is a modified version of ResNet-50 [12] (following [31], the first 7x7 convolution kernel is replaced with a 3x3 kernel and the first max pooling layer is removed), while the unmodified version is used in WIDER Attribute and CUB-200-2011. Each network is pretrained for 1000 epochs using the momentum SGD optimizer. Other hyperparameter settings are provided in the supplementary material. For CLRC, the dimension d_1 in the bottleneck module is fixed at 1024 for all datasets. The dimension d_2 is set to 128 for UT Zappos 50k and WIDER Attribute, and 64 for CUB-200-2011. For simplicity, the output size of the group-specific linear projectors is set to half of the dimension of d_2. The quality of the learnt representation is validated via the linear evaluation protocol, where a randomly initialized linear layer is attached to the frozen pretrained encoder and fine-tuned on the downstream object class labels.

Results. Tab. 2 compares the top-1 and top-5 accuracy on the visual attribute datasets. Interestingly, the previously overlooked cross-entropy baseline performs competitively across all three datasets, and even achieves the best performance in UT Zappos 50k. On the other hand, the proposed CLRC approach attains the second-best top-1 accuracy in UT Zappos 50k and the best overall performance in WIDER Attribute and CUB-200-2011. Notably, there is a significant performance gap between CLRC and the second-best methods in CUB-200-2011 (4.88% and 5.55% gap in top-1 and top-5 accuracy, respectively). Compared to

Method	UT Zappos 50k [38]		WIDER Attribute [19]		CUB-200-2011 [35]	
	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.
SimCLR [4]	82.18±0.05	99.11±0.01	39.28±0.10	69.27±0.14	24.79±0.23	52.49±0.19
BYOL [8]	84.59±0.08	98.46±0.16	44.29±0.22	73.12±0.29	28.71±0.23	55.38±0.17
SwAV [2]	80.15±0.10	98.56±0.11	38.72±0.26	69.31±0.23	14.24±0.20	33.18±0.23
SupCon [16]	*88.55±0.03*	*99.57±0.00*	*46.78±0.07*	*73.94±0.04*	*59.94±0.16*	*82.27±0.20*
Cross-entropy	**86.27±0.02**	**99.68±0.03**	37.41±0.22	67.31±0.20	40.39±0.42	66.65±0.17
CMC [30]	85.75±0.11	99.66±0.02	40.14±0.20	69.70±0.09	37.25±0.22	66.86±0.20
Cl-InfoNCE [31]	85.19±0.05	99.51±0.02	43.74±0.12	73.76±0.08	26.48±0.16	52.58±0.02
CCL-K [32]	85.35±0.07	99.59±0.03	8.80±0.09*	30.74±0.16*	34.24±0.17	64.78±0.13
CLRC (this work)	86.03±0.02	99.37±0.00	**45.37±0.48**	**74.08±0.13**	**45.27±0.12**	**72.41±0.07**

Table 2. Top-1 and Top-5 classification accuracy (%) on the testing set of each dataset. The self-supervised methods (SimCLR, BYOL, SwAV) and supervised contrastive learning on downstream labels (SupCon) represent the performance lower bounds and upper bound, respectively. * is unable to converge, despite our best effort in adapting CCL-K to this dataset.

other datasets, CUB-200-2011 contains the most number of binary attributes (up to 312 distributed across 28 semantic groups), suggesting that CLRC is very effective in scaling to a large set of attributes. In terms of parameter efficiency, the number of learnable parameters in the auxiliary modules (i.e., modules excluding the base encoder) of CLRC for UT Zappos 50k, WIDER Attribute and CUB-200-2011 is 2.9M, 4.1M, and 4M, respectively. This is in contrast to the common MLP projector design (a three-layer MLP with the configuration 2048-2048-128) used in the other contrastive learning baselines, which consists of 8.7M learnable parameters. This illustrates the parameter efficiency of the proposed group-wise bottleneck module and its flexibility in scaling to different number of attribute groups.

4.3. Adaptation to Fair Representation Learning

Dataset. When deploying deep learning models, an important consideration is how the models will behave towards different groups of people, especially those who are underrepresented in the training data. This has led to the study of fair representation learning, which aims to reduce the disparity among different subgroups [21]. In this section, an adaption of CLRC in the context of fair representation learning is demonstrated. Using the Fitzpatrick17k [9] dataset as an example, we propose to treat the sensitive skin type attributes as individual semantic groups. Fitzpatrick17k consists of skin dermatology images annotated with skin type and lesion labels. Following prior work [42], the lesion labels are converted into a binary label ("*benign*" vs. "*malignant*"). To adapt CLRC to Fitzpatrick17k, the global representation is projected into six group-specific feature vectors, where each group is assigned to a skin type in the dataset. Within each group, the binary lesion label is represented as a two-dimensional one-hot attribute vector. A third attribute "*ignored*" is added to differentiate images that does not belong to the assigned group, *e.g.*, if the image does not belong to the assigned group, its lesion label is ignored and its one-hot vector will look like $[0, 0, 1]$. During group-wise contrastive learning, two images are considered as a positive pair if they have the same attribute label ("*benign*", "*malignant*", or "*ignored*"). This encourages each group to be specialized in learning subgroup-specific features, and thereby prevent underrepresented subgroups from being overwhelmed by the majority subgroups.

Experiment Setup. Following the protocol introduced in MEDFAIR [42], the Fitzpatrick17k dataset is divided into training/validation/testing sets with a proportion of 80/10/10. The base encoder is a ResNet-18 initialized with weights pretrained on the ImageNet-1k [27] dataset. For the baseline, the base encoder is fine-tuned on the training set using the momentum SGD optimizer with a batch size of 1024. Training is stopped if the validation worst-case AUC-ROC (i.e., AUC-ROC on the lowest performing subgroup) does not improve for five epochs. In the rest of the experiments, the base encoder is first pretrained for 50 epochs using different pretraining objectives before the fine-tuning step. For the bottleneck module in CLRC, the output size of g_1 and g_2 is set to 512 and 128, respectively.

Results. Adopting the same evaluation metrics as MEDFAIR [42], the fairness of the learnt representation is measured against three metrics: overall AUC-ROC (higher is better), worst-case AUC-ROC (higher is better) and the gap between the best- and worst-case AUC-ROC (lower is better). Tab. 3 shows the results on the testing set of Fitzpatrick17k. SimCLR and SupCon substantially improve the

overall AUC-ROC but they significantly reduce the worst-case performance and widen the gap between different skin types. In other words, pretraining without considering the sensitive attributes will negatively impact the performance in underrepresented subgroups. Methods that are aware to the sensitive attributes such as CMC and CLRC are able to improve the subgroup performance. In particular, CLRC significantly outperforms the other methods in terms of worst-case AUC and AUC gap, showing its potential to learn fairer representations.

Method	Overall (↑)	Min. (↑)	Gap (↓)
Baseline	88.16	79.25	13.75
SimCLR [4]	89.25 (+1.09%)	74.84 (-4.41%)	20.72 (+6.97%)
SupCon [16]	91.67 (**+3.51%**)	74.84 (-4.41%)	21.64 (+7.89%)
CMC [30]	87.59 (-0.57%)	80.29 (+1.04%)	12.63 (-1.12%)
CLRC (this work)	88.65 (+0.49%)	84.28 (**+5.03%**)	8.33 (**-5.42%**)

Table 3. AUC-ROC (%) on the testing set of Fitzpatrick17k [9]. The baseline is a ImageNet initialized ResNet-18 fine-tuned on the binary lesion labels. Changes relative to the baseline are indicated in the parentheses. (Overall: overall AUC-ROC across every skin type, Min.: worst-case AUC-ROC, Gap: gap between the best- and worst-case AUC-ROC)

4.4. Adaptation to Hierarchical Labels

Dataset. Hierarchical labels encode a subsumptive relationship between different labels and are usually represented as a tree or a directed acyclic graph. In the ImageNet [6] dataset, images are organized according to the WordNet [22] hierarchy, where non-leaf nodes represent coarse-grained labels of varying granularity, while leaf nodes represent fine-grained object classes. To adapt hierarchical labels to CLRC, each level of the hierarchy could be interpreted as its own group, and the group-wise contrastive learning objective becomes a hierarchy-wise objective. As a demonstration, the efficacy of CLRC in learning from hierarchical labels is evaluated on ImageNet-100, a subset of ImageNet-1k dataset [27] which contains 100 fine-grained classes. For this experiment, the coarse-grained labels from the 5-th and 6-th level of the WordNet hierarchy (the fine-grained classes is at the 14-th level) are extracted, forming two groups of coarse labels with a total of 13 classes.

Experiment Setup. The experiment setup is similar to the setup described in Sec. 4.2 (more details are provided in the supplementary materials). ResNet-50 encoders are pretrained on the coarse-grained labels before fine-tuning on the fine-grained labels via the linear evaluation protocol. For CLRC, the dimensions of g_1 and g_2 are respectively 2048 and 128.

Results. The results of hierarchical representation learning are given in Tab. 4. The results include an additional baseline, HiMulConE [40], which is proposed specifically to utilize the hierarchical relationship in labels. Although HiMulConE also performs contrastive learning in a hierarchy-wise manner, there are two main differences with regard to how the contrastive loss is implemented between HiMulConE and the proposed adaption of CLRC to hierarchical labels: 1) HiMulConE enforces a hierarchy constraint so that the losses towards the deeper levels are larger than those at the shallower levels by weighting the losses based on the depth of the tree; CLRC does not incorporate any hierarchical constraint and instead rely on the hierarchy-specific entropies to weight the losses, 2) HiMulConE uses a single MLP projector across all hierarchies while CLRC uses a separate light-weight linear projector for each hierarchy. CCL-K is excluded from this experiment as it uses continuous labels extracted from a pretrained vision-language model instead of hierarchical labels. In terms of Top-1 and Top-5 accuracy, CLRC achieves the best performance, with slight improvements over Cl-InfoNCE and HiMulConE. The results suggest that CLRC is a very competitive method to learn from hierarchical labels.

Method	Top-1 Acc.	Top-5 Acc.
SimCLR [4]	66.98	87.99
BYOL [8]	70.02	87.33
SupCon [16]	87.11	96.29
Cross-entropy	69.52	88.15
CMC [30]	74.52	91.08
Cl-InfoNCE [31]	75.29	91.90
HiMulConE [40]	<u>76.14</u>	<u>91.73</u>
CLRC (this work)	**76.65**	**92.09**

Table 4. Top-1 and Top-5 classification accuracy (%) on the testing split of ImageNet-100. SimCLR and BYOL represent the performance lower bounds while SupCon represents the performance upper bound.

4.5. Ablation Study

Ablation studies are conducted on the ImageNet-100 benchmark. The results are tabulated in Tab. 5.

Ablation	Top-1 Acc.	Top-5 Acc.
Dimension d_1 = 1024	75.49	91.69
Dimension d_1 = 2048*	76.65	**92.09**
Dimension d_1 = 4096	**77.44**	92.04
Dimension d_2 = 64	76.07	92.37
Dimension d_2 = 128*	76.65	92.09
Dimension d_2 = 256	**77.38**	**92.4**
Dimension d_2 = 512	77.12	91.92
shared projector	76.56	91.94
group-specific projectors*	**76.65**	**92.09**
uniform weighting in $\mathcal{L}_{cl}$	75.51	91.54
entropy weighting in $\mathcal{L}_{cl}$*	**76.65**	**92.09**
$\mathcal{L}_{cl}$ only (remove $\mathcal{L}_{mse}$)	75.84	91.65

Table 5. Results of ablation study on CLRC in the ImageNet-100 benchmark. * is the default configuration.

Dimension of d_1. $g1$ maps the global representation into an intermediate representation with dimension d_1, and is analogous to the first layer of the MLP projector used in prior contrastive learning approaches [4, 11]. The commonly adopted output size is 2048, which matches the dimension of the representation extracted from a ResNet-50 encoder. Here, a lower dimension (1024) was tested and was shown to perform worse than the default value of 2048.

Dimension of d_2. d_2 controls the size of each group-specific representation where the group-wise contrastive learning objective is applied. A larger dimension (256/512) was able to improve the representation quality while a smaller dimension (64) was still able to yield a competitive performance.

Shared vs. Separate Linear Projectors. The use of a shared linear projector (to project the disentangled group-specific features) across all semantic groups slightly reduces the Top-1 and Top-5 accuracy, which indicates that a group-specific projector is not as important. Combined with the study on the dimension of d_2, the results suggests that majority of the improvement in representation quality comes from the disentanglement layers and the group-specific contrastive learning objective. Thus, a shared linear projector might be a viable design to save memory cost.

Uniform vs. Entropy-weighted Contrastive Loss. Uniform weighting assigns an equal weight to each group-specific contrastive loss ($\mathcal{L}_{gcl}$). As demonstrated in Tab. 5, this results in a degradation of the representation quality. On the other hand, the proposed entropy-weighting scheme provides an effective and tuning-free weighting method for estimating the importance of each loss term through the group entropies computed from the training data.

Removing Reconstruction Loss. As shown in the last row of Tab. 5, removing the reconstruction loss term ($\mathcal{L}_{mse}$) causes performance drop in both Top-1 and Top-5 accuracy. This shows that the objective of maximizing the consistency between the reconstructed representation and the representation from another augmented view can provide additional training signals which encourage the learnt representations to be robust to different augmentations.

5. Discussion

Through empirical experiments on various types of weak labels, CLRC was shown to be very effective in learning good representations under the weakly supervised setting, assuming that the labels are discrete and can be grouped into semantically related categories. One current limitation of the proposed method is that it cannot be applied directly to settings involving continuous labels such as text embeddings. Thus, an investigation on how to adapt the grouping and loss weighting mechanism of CLRC to the continuous setting would be a promising direction for future work. Another interesting but orthogonal research direction would be to integrate the weakly-supervised pretraining framework into the task of compositional zero-shot learning [28], where the learnt representations could potentially provide additional benefits in terms of convergence speed and downstream performance.

6. Conclusion

A pretraining framework for weakly supervised representation learning is proposed in this work. Specifically, a group-wise contrastive bottleneck module is presented to leverage the semantic grouping information in the label space. Two learning objectives consisting of a group-wise contrastive loss and a reconstruction loss are introduced to learn robust and transferable representations. Compared to prior approaches, CLRC is more parameter-efficient and is more flexible in adapting to various types of labels, including visual attributes, sensitive attributes and hierarchical labels. In terms of downstream performance, CLRC outperforms prior approaches in most benchmark datasets, further closing the gap with fully supervised pretraining.

Acknowledgements

The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924, 2020. 5, 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9630–9640, 2021. 1

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 1, 3, 4, 5, 6, 7, 8

[5] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1635–1643, 2015. 3

[6] Jia Deng, W. Dong, R. Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2009. 5, 7

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020. 1, 3, 5, 6, 7

[9] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021. 5, 6, 7

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15979–15988, 2021. 1

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735. IEEE, 2020. 1, 3, 4, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016. 5

[13] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021. 3

[14] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7014–7023. IEEE, 2018. 3

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. pmlr, 2015. 4

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 3, 5, 6, 7

[17] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016. 3

[18] Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991. 2

[19] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, 2016. 5, 6

[20] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision*, pages 181–196, 2018. 1, 3

[21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Ani Saxena, Kristina Lerman, and A. G. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54:1 – 35, 2019. 6

[22] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 7

[23] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. Focus on the positives: Self-supervised learning

for biodiversity monitoring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2, 3

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[25] Jizong Peng, Ping Wang, Christian Desrosiers, and Marco Pedersoli. Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021. 3

[26] Martin Rajchl, M. J. Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Bernhard Kainz, and Daniel Rueckert. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Transactions on Medical Imaging*, 36:674–683, 2017. 3

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6, 7

[28] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, June 2022. 8

[29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1195–1204, 2017. 4

[30] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, 2020. 5, 6, 7

[31] Yao-Hung Hubert Tsai, Tianqin Li, Weixin Liu, Peiyuan Liao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning weakly-supervised contrastive representations. In *International Conference on Learning Representations*, 2022. 2, 3, 5, 6, 7

[32] Yao-Hung Hubert Tsai, Tianqin Li, Martin Q. Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Conditional contrastive learning with kernel. In *International Conference on Learning Representations*, 2022. 2, 3, 5, 6

[33] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 4

[34] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 4

[35] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5, 6

[36] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284. IEEE, 2020. 3

[37] Wei Xia, Csaba Domokos, Jian Dong, Loong Fah Cheong, and Shuicheng Yan. Semantic segmentation without annotating segments. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2176–2183, 2013. 3

[38] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014. 5, 6

[39] Jianfu Zhang, Yuanyuan Huang, Yaoyi Li, Weijie Zhao, and Liqing Zhang. Multi-attribute transfer via disentangled representation. In *AAAI Conference on Artificial Intelligence*, 2019. 3

[40] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16639–16648, 2022. 2, 3, 7

[41] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xiansheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019. 3

[42] Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking fairness for medical imaging. In *International Conference on Learning Representations*, 2023. 6