

Benchmarking Omni-Vision Representation through the Lens of Visual Realms

Yuanhan Zhang¹, Zhenfei Yin², Jing Shao², and Ziwei Liu¹

¹ S-Lab, Nanyang Technological University, Singapore

{yuanhan002, ziwei.liu}@ntu.edu.sg

² SenseTime Research

{yinzhengfei, shaojing}@sensetime.com

Abstract. Though impressive performance has been achieved in specific visual realms (*e.g.* faces, dogs, and places), an omni-vision representation generalizing to many natural visual domains is highly desirable. But, existing benchmarks are biased and inefficient to evaluate the omni-vision representation—these benchmarks either only include several specific realms, or cover most realms at the expense of subsuming numerous datasets that have extensive realm overlapping. In this paper, we propose Omni-Realm Benchmark (**OmniBenchmark**). It includes 21 realm-wise datasets with 7,372 concepts and 1,074,346 images. Without semantic overlapping, these datasets cover most visual realms comprehensively and meanwhile efficiently. In addition, we propose a new supervised contrastive learning framework, namely **Relational Contrastive learning (ReCo)**, for a better omni-vision representation. Beyond pulling two instances from the same concept closer—the typical supervised contrastive learning framework—ReCo also pulls two instances from the same semantic realm closer, encoding the semantic relation between concepts, facilitating omni-vision representation learning. We benchmark ReCo and other advances in omni-vision representation studies that are different in architectures (from CNNs to transformers) and in learning paradigms (from supervised learning to self-supervised learning) on OmniBenchmark. We illustrate the superior of ReCo to other supervised contrastive learning methods, and reveal multiple practical observations to facilitate future research. The code and models are available at <https://zhangyuanhan-ai.github.io/OmniBenchmark>.

Keywords: Representation Learning, Visual Realm

1 Introduction

Large-scale pre-trained models, either trained in a supervised [24, 62, 66] or unsupervised manner [7, 23, 40], have become a foundation [4] for modern computer vision [9, 19, 28, 29, 40, 42, 47, 68]. The generalization quality of pre-trained models—whether models are helpful for various downstream tasks—typically determines models’ quality. Recent benchmarks for evaluating model generalization mainly collect downstream tasks belonging to two facets: across image

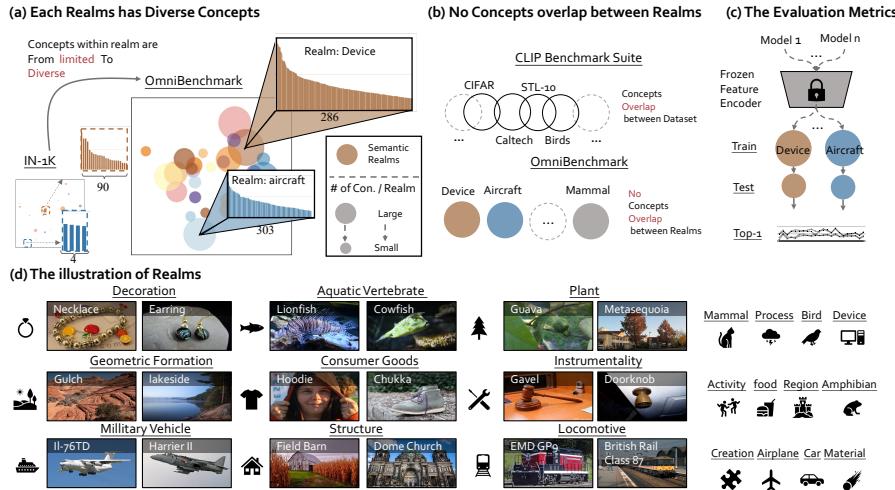


Fig. 1. A overview of OmniBenchmark. (a) Compared with ImageNet-1k, OmniBenchmark covers more realms and annotate more images of each realm. (b) Included datasets in OmniBenchmark have no semantic concepts overlap. (c) OmniBenchmark diagnoses pre-trained models through linear probing. (d) We present 21 semantic realm-wise datasets and the sampled concepts of 9 datasets.

domains (*e.g.* from natural domain to synthetic domain) [68]; across vision tasks (*e.g.* from image classification to instance segmentation) [7, 23]. In this study, we focus on the third facet: across semantic super-concepts/realms (*e.g.* across pets to scenes). We found that current benchmarks focusing on the last facet either cover a limited range of the semantic realms or are overly cumbersome for evaluation. For example, ImageNet-1k [43] only focuses on the mammal, instrumentality, devices, and consumer goods. But these realms struggle to describe the complete realms in the natural domain. Meanwhile, though CLIP [40] builds a significant benchmark that consists of 24 datasets across a large spectrum of realms, several datasets included have an extensive concept overlapping (*e.g.* CIFAR10 [31], Caltech101 [18], and STL-10 [12] all have bird class) as shown in Fig. 1(b), resulting in benchmarking on its benchmark suite is less efficient than on an ideal benchmark where datasets included do not conceptually overlap with each other.

In this work, we systematically investigate how to build a benchmark for evaluating omni-vision representation that can generalize to a wide range of semantic realms. This benchmark focuses the classification task on the natural image domain. Through the analysis of ImageNet-1k, we find that the limited number of concepts in its concepts resource, *i.e.* WordNet [37], results in its limited realm coverage.¹ Starting from WordNet, we enrich its concepts by integrating new concepts from Wikidata [54], building a large ontology with nearly

¹ Annotation budget also limits its realm coverage.

100,000 concepts. These concepts illustrate a more complete distribution of semantic realms. Further, we separate these concepts into 21 realm-wise datasets, ensuring these datasets have no overlapping semantic concepts. After carefully annotation, we build a benchmark consisting of 7,372 classes and 1,074,346 Creative Commons licenses (CC-BY) data across 21 realms. We illustrated the construction of our benchmark in Fig. 1, and we refer to it as the Omni-Realm benchmark (**OmniBenchmark**).² OmniBenchmark has two appealing properties. **1) Diversity.** the number of realms of OmniBenchmark is twice of ImageNet-1k’s, *i.e.* 21 *v.s.* 9 , and the average number of concepts per realm is nine times bigger than ImageNet-1k, *i.e.* 263 *v.s.* 29. **2) Conciseness.** Since realm-wise datasets have no concept overlapping, OmniBenchmark is concise.

We further investigate how to learn a better omni-vision representation. We purpose an omni-vision representation should not only cluster instances of the *same* concept but also cluster the instances of *related* semantic concept—two concepts within the same realm should imply closer relation than two concepts across the realm, *e.g.* the relation between husky and labrador (two dog species) should be closer than husky and Ferrari 488 (a sports car). However, current representative representation learning methods, *e.g.* supervised contrastive learning [13, 26], commonly construct negative pairs by exhaustive sampling without considering their semantic relation, *e.g.* husky and labrador should have the same possibility of being the negative pair than husky and Ferrari 488. Motivated by this limitation, we present a novel supervised contrastive learning framework called **Relational Contrastive Learning (ReCo)**. ReCo selects the high-quality negative pairs, which belong to different semantic realms, for supervised contrastive learning. ReCo improves the state-of-the-art (SOTA) supervised contrastive learning method (PaCo [13]) on the ImageNet-1k by 0.7 points gain.

We conduct extensive studies on OmniBenchmark to diagnose ReCo and other advances in the omni-vision representation learning including architectures (from CNNs [24] to transformers [15]); learning paradigms (from fully-supervised learning to self-supervised learning); pre-training data volume (from ImageNet-1k [14] to IG-1B [62]). We reveal several valuable observations and prove the priority of ReCo: ReCo outperforms PaCo with an average 0.5 points gain on OmniBenchmark.

We summarize the contributions of this work as follows.

- We propose OmniBenchmark with 21 semantic realm-wise datasets. OmniBenchmark focus on evaluating the concept generalization ability of omni-vision representation thoroughly and efficiently.
- We evaluate 22 recent representation learning approaches on OmniBenchmark, uncovering several interesting insights for future research.
- A novel supervised contrastive learning method, ReCo, is proposed for encoding the semantic relation information in the supervised contrastive learning framework, achieving competitive performance on ImageNet1k and OmniBenchmark.

² We use “Omni” to emphasize the diversity of semantic realms.

2 Related Work

Representation Learning Methods. Representation learning has advanced thanks to improvements in various learning paradigms. To avoid the need for supervision, self-supervised learning [6, 7, 10, 16, 23, 32, 34, 38, 58, 69] has been proposed. Recently, Guo *et al* [21] proposed hierarchical contrastive selective coding (HCSC) that improves conventional contrastive learning by implicitly coding the semantic similarity between instances. Inspired by HCSC, ReCo explicitly encodes the semantic similarity, using the label information in the hierarchical concepts structure. In addition, weakly supervised learning [61, 62, 75] focuses on learning from unlabeled data by self-training. The success of representation learning should also owe to architecture designs breakthrough, *e.g.* Vision Transformer (ViT) [15]. This paper benchmark the most recent methods that have public implementations available. Moreover, there have extensive studies that explore to use semantic structure in various ways. Specifically, Bertinetto *et al* [3] use the information in the class hierarchy to achieve competitive performance on several datasets. Wang *et al* [57] model the classification process on the semantic hierarchy as a sequential decision-making task. In addition, Wang *et al* [56] propose a deep fuzzy tree model for learning the semantic hierarchy better. Wang *et al* [55] and Guo *et al* [20] propose very insightful frameworks for leveraging semantic relation information in tasks other than image classification.

Evaluations and Benchmarks. The importance of empirical evaluation of representation learning is highlighted by the growing number of major evaluation papers [9, 19, 29, 42, 47]. CLIP [40] proposes a significant benchmark that consists of 24 datasets across different semantic realms. However, CLIP benchmark suite struggles to select the most valuable set of datasets for the benchmark. In particular, several datasets have an extensive concept overlapping, making their benchmark suite cumbersome for evaluation. Visual Decathlon [41] evaluates the ability of representations to capture simultaneously ten very different visual domains and measures their ability to perform well uniformly. Visual Task Adaptation Benchmark (VTAB) [68] includes datasets from several different domains (natural, specialized, and structured) and annotation information (classification, counting, and *etc*) to evaluate methods of varying learning paradigms. We argue that the motivation and evaluation protocol of OmniBenchmark are different from VTAB and Visual Decathlon. First, the Visual Decathlon studies the multi-task learning ability of models. Its evaluation protocol is train-test. By contrast, OmniBenchmark explores the generalization ability of pre-trained models. Its evaluation protocol is pretrain-train-test. Secondly, VTAB quantifies the generalization ability of representation to transfer to different image domains and different annotation information (The first and second facet of quantifying representation quality as mentioned in Sec.1). OmniBenchmark focuses explicitly on evaluating concept generalization of the classification task in the natural domain. Beyond that, ImageNet-COG [44] also studies concept generalization. It evaluates the generalization of models from ImageNet-1k to other unseen concepts for ImageNet-1k. OmniBenchmark evaluates the generalization of models

from any open pre-trained source to extensive categories organized by semantic realms.

3 The Construction of Omni-Realm Bechmark

In this section, we describe the construction of OmniBenchmark. First, we integrate new concepts from Wikidata into WordNet, enlarging the concept storage of WordNet extensively (Sec. 3.1). For the all concepts of this larger “WordNet”, we secondly crawl raw images of each of them from Flickr (Sec. 3.1). Thirdly, we select valid concepts from the whole concepts, following multiple steps (Sec. 3.2). Fourthly we split WordNet into 21 sub-trees, each sub-tree represents a semantic realm and covers a set of valid concepts. (Sec. 3.3) Fifthly, we annotate all the raw images of the valid concepts of these 21 semantic realms (Sec. 3.4), forming the 21 realm-wise datasets for benchmarking. In addition, we present detail information of the statistics (Sec. 3.5), and the evaluation protocols of OmniBenchmark. (Sec. 3.6).

3.1 Integrating New Concepts

Wikidata [64] contains a large number of concepts, such as different kinds of foods and structures. As the number of concepts in Wikidata continues to grow, we have integrated 170,586 concepts from it so far. These concepts are the leaf nodes in the taxonomy of the Wikidata. Referred to [50], we link Wikidata leaf node concepts to the WordNet through leveraging on the “sub-classOf” attribute of Wikidata. This attribute denotes the hypernyms relation. Further, we obtain raw images for each concept by using the Flickr API.³

3.2 Concept Filtering

After integrating 170,586 new concepts into WordNet, there are nearly 210K concepts in the WordNet. However, not all these concepts are valid for the image classification benchmark. Thus, we manually filter the 210K concepts by the following steps. 1) We ask annotators to identify and then discard concepts that are related to the offensive content, such as drugs and blood. 2) Referred to Yang *et al* [63], we only keep the visual concepts that are highly concrete words. 3) Referred to ImageNet [43], We suggest ensuring the label in our benchmark is mutually exclusive. We thus only keep annotate leaf node in the larger “WordNet” created in Sec. 3.1. 4) In our trial annotation, *i.e.* annotation for randomly selected 200 concepts, only 50% of the crawled data of a specific concept are semantic related to this concept. Since we plan to have 50 images for the test-set of the concept, and 50 images for its train set, we thus discard concepts that have less than 200 crawled data. We illustrate these steps in Fig. 2 (b).

³ <https://www.flickr.com/>, All the crawled images are strictly followed the CC-BY license.

3.3 Realm Selection

In WordNet, concepts are organized by the hierarchical structure with many sub-trees. We select semantic realms from these sub-trees by following three principles. 1) We select sub-trees that cover at least 20 valid concepts. 2) We discard the sub-tree that is covered by another sub-tree. 3) We discard the sub-tree that most of their concepts imply non-natural images and personal information. For example, we discard the chemical material realm because most of its concepts can only be related to the structural formula image. In addition, we also discard person because its related images contain personal information such as IDs, names. We illustrate these three principles in the Fig. 2 (c).

3.4 Image Annotation and De-Duplication

For each candidate raw image, we ask five annotators whether this image conforms to its query concept. An image is annotated only if at least 3 out of 5 annotators consider its semantic information closely related to its query concept. More detail information, *e.g.* the annotation interface, is described in the *Supplementary Material*.

To enable a meaningful test of generalization, OmniBenchmark remove duplicates with potential pre-training datasets that includes Bamboo-CLS [45, 70], ImageNet-22K [14], PASCAL-VOC [17], MS-COCO [35] and Places [71]. Specifically, we firstly utilize Difference Hash (DHash) [2] to calculate the hash-code of images of both OmniBenchmark and these pre-training datasets. Then we delete the image that has the same hash-code as any image in these pre-training datasets.

3.5 Benchmark Statistics

In Fig. 2 (a), we compare the distribution of the concepts of each realm of OmniBenchmark and ImageNet-1k. Specifically, 815 concepts of ImageNet-1k are included in these 21 realms, which can fairly reflect the distribution of the ImageNet-1k.⁴ Fig. 2 (a) presents that ImageNet-1k covers a very limited number of *ImageNet-1k-seen-realms*, *e.g.* Device, instrumentality consumer_goods and mammal.⁵ Therefore, ImageNet-1k hardly becomes an ideal benchmark to evaluate the generalization ability of the omni-vision representation. In contrast, the number of OmniBenchmark-seen-realms is 21 that is twice of ImageNet-1k-seen-realms'. We believe the large spectrum of realms included in OmniBenchmark can relatively thoroughly represents the distribution of realms in the natural domain.

We carefully discuss the potential legal issue of OmniBenchmark, *e.g.* the copyright and privacy issue, in the *supplementary material*.

⁴ The other 185 concepts are included in realm that are filtered in Sec. 3.3.

⁵ In the following material, the dataset-seen-realms/dataset-unseen-realms is a set of realms. Each dataset-seen-realm/dataset-unseen-realm in it has at least 20 concepts/fewer than 20 concepts.

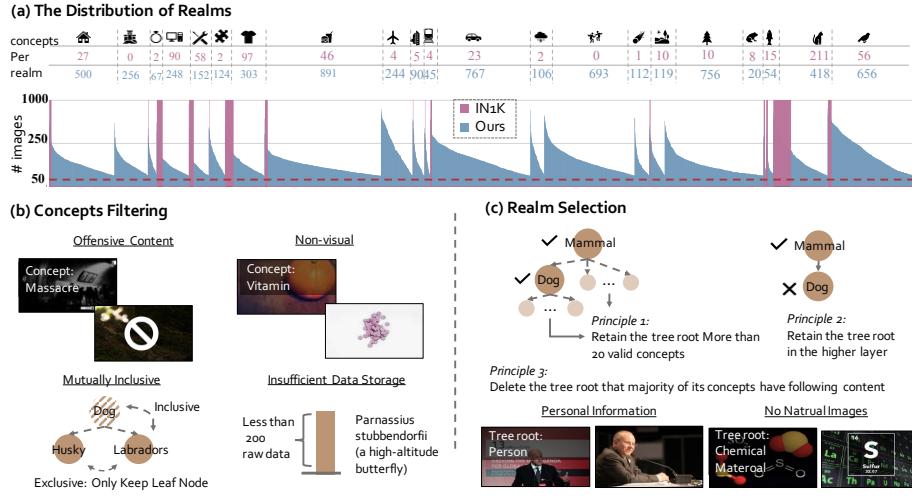


Fig. 2. (a) The statistic of OmniBenchmark. We compare the distribution of the concepts of each realm in OmniBenchmark and ImageNet-1k. For ImageNet-1k, 12 out of 21 realms are unseen-realm because ImageNet-1k has less than 20 concepts in these realms. We report the number of concepts per realm of ImageNet-1k and OmniBenchmark in a different color. **(b) Four principles of filtering concepts.** We discard the concept when it implies offensive content, is a non-visual word, is not a leaf node, and has less than 200 raw data. **(c) Three principles of selecting realms.** We illustrate three principles to select semantic realms from numerous sub-trees.

3.6 Evaluation Protocol

We now present the protocol for OmniBenchmark, and summarize the metrics for the experiments presented in Sec. 5.

Linear Probing. Our benchmark quantifies the generalization ability of visual representation on different semantic realms. We build the evaluation protocol based on the assumption that a good representation is an omni-vision representation that can generalize any specific realm without updating the feature extractor. Therefore, we keep the pre-trained backbone frozen and use it as the feature extractor. Then, we learn linear logistic regression classifiers for realm-wise datasets, following the linear probing setting [19].

Metrics. We report the top-1 accuracy for all experiments. In addition, to make the plots clearer and the differences easier to grasp, we follow the metric setting in the ImageNet-CoG [44] that plots accuracy relative to the baseline ResNet-50 model pre-trained on ImageNet-1k.

4 Methodology

4.1 Preliminaries

Self-supervised Contrastive learning. A widely-used way to achieve self-supervised learning is contrastive learning, which is proposed in InfoNCE [51]. InfoNCE loss works for pulling positive pairs belonging to the same classes closer and push negative pairs belonging to different classes away is an effective way to achieve self-supervised learning. Specifically, for a set of N sampled pairs, $\{x_k, y_k\}_{k=1\dots N}$, the corresponding batch used for training consists of $2N$ pairs, $\{\tilde{x}_i, \tilde{y}_i\}_{i=1\dots 2N}$, where \tilde{x}_{2k} and \tilde{x}_{2k-1} are two random augmentations of x_k and $\tilde{y}_{2k} = \tilde{y}_{2k-1} = y_k$. Finally, a N samples batch forms a $2N$ multiview batch, let $i \in I \equiv \{1\dots 2N\}$ be the index of an augmented sample, and let $j(i)$ be the index of the other augmented sample originating from the same source sample. For a given sample i , the loss takes the following form:

$$\mathcal{L}_i = -\log \frac{\exp(z_i \cdot z_{j(i)})}{\sum_{a \in A(i)} \exp(z_i \cdot z_a)}. \quad (1)$$

Here, $z_i = Encoder(x_i)$, $A(i) \equiv I \setminus \{i\}$, the index i is called the anchor, index $j(i)$ is called the positive, and the other $2(N - 1)$ indexes ($A(i) \equiv I \setminus \{i\}$) are called the negative.

Supervised Contrastive learning. To add label information into the self-supervised representation learning, Khosla *et al* [26] proposed the supervised contrastive learning (Supcon) in the following way:

$$\mathcal{L}_i = -\sum_{z_+ \in P(i)} \log \frac{\exp(z_i \cdot z_+)}{\sum_{z_k \in A(i)} \exp(z_i \cdot z_k)}, \quad P(i) \equiv \{p \in A(i) : \tilde{y}_p = \tilde{y}_i\}. \quad (2)$$

Parametric Contrastive learning. Cui *et al* [13] introduce a set of parametric class-wise learnable center $\mathbf{C} = \{c_1, c_2, \dots, c_n\}$ into the original supervised contrastive learning, and named a new framework: Parametric Contrastive learning (PaCo). Here, n is the number of the classes. Specifically, the loss is change to.

$$\mathcal{L}_i = -\sum_{z_+ \in P(i) \cup \{c_{\tilde{y}}\}} \log \frac{\exp(z_i \cdot z_+)}{\sum_{z_k \in A(i) \cup \mathbf{C}} \exp(z_i \cdot z_k)}. \quad (3)$$

PaCo boosts the performance of Supcon on several datasets.

4.2 Motivation

Previous Supcon and PaCo methods sample negative samples uniformly over the datasets. However, some negative samples are semantics closely to the query sample, *e.g.* the husky and labrador in the Fig 3 (a). Wrongly expelling these negative samples from query samples could break the semantic structure to some extent, hampering the representation learning.

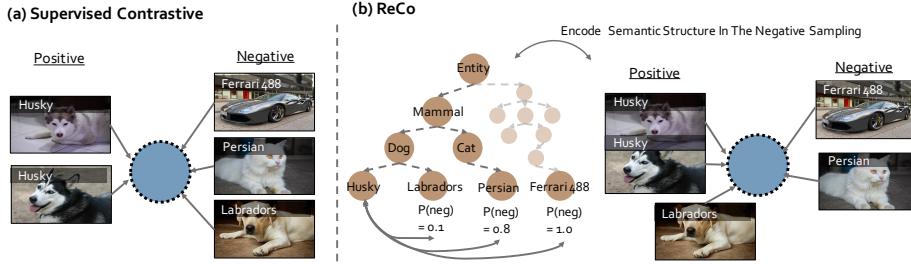


Fig. 3. Supervised contrastive losses *v.s* ReCo. The supervised contrastive loss (left, Eq. 2) contrasts the set of all samples from the same class as positives against the negatives from the remainder of the batch. The ReCo (right, Eq. 6), however, contrasts the set of all samples from the same class as positives against the *real semantic dissimilar* negatives from the remainder of the batch. Leverage semantic relation when sampling negative samples, ReCo provides an embedding space where elements of similar classes (husky and labrador) are aligned closer than in the supervised case.

Motivated by this limitation, we aim to carefully select semantic irrelevant negative samples to the query sample instead of selecting negative samples uniformly. In particular, We propose a novel hierarchical instance contrastive learning framework called Relational Contrastive Learning (ReCo). ReCo captures the semantic relation and then transforms this relation into the probability of being sampled as negative samples, as shown in Fig. 3 (b).

4.3 Relational Contrastive learning

The gist of ReCo is to pulling instances of similar semantic classes closer while pushing instances of dissimilar semantic classes far apart. For a specific query image x , we select real semantic dissimilar negative samples for the contrastive learning by performing Bernoulli sampling on each negative candidate, which is inspired by HCSC [21]. To achieve this goal, we first define a similarity $s(m, n)$ measure between two classes m, n as follows.

$$s(m, n) = -\log \frac{d_{\min}(m, n) + 1}{2 * \max(l^m, l^n) + 1}, \quad (4)$$

m, n lie in the hierarchical depth l^m, l^n , and $d_{\min}(m, n)$ denotes the shortest path that connects the m and n in the hypernym/hyponym taxonomy. As shown in Fig 3 (b), since the dog is the father node of the husky, their shortest distance equals one, and the shortest distance between husky and labrador equals two because they are in sibling relation. In addition, $\max(\cdot)$ in the denominator makes the $s(m, q) > s(m, n)$ if q is another class that lies in deeper depth than n and $d_{\min}(m, n) = d_{\min}(m, q)$. We ensure that m should be more related to node that lies in deeper depth because as the hierarchical depth is deeper, the nodes in that depth are more concrete. For example, as shown in Fig 3 (b), the

dog should be semantic closer to the husky than to the mammal. Moreover, we normalize the similarity between m and other classes for ensuring $s(m, \cdot) \in [0, 1]$.

On such bases, we conduct negative sample selection for a specific instance x_i . For the negative candidate $z_k \in A(i)$, we are more likely to select it if its $S(\tilde{y}_i, \tilde{y}_k)$ is low. The negative sampling follows the Bernoulli sampling:

$$A_{select}(i) = \{\beta(z_k; P = 1 - s(\tilde{y}_i, \tilde{y}_k)) | z_k \in A(i)\}, \quad (5)$$

where $\beta(z; s)$ denotes a Bernoulli trial of accepting z with similarity S . By using these refined negative samples selection, we define the objective function of ReCo as below

$$\mathcal{L}_i(ReCo) = - \sum_{z_+ \in P(i)} \log \frac{\exp(z_i \cdot z_+)}{\sum_{z_k \in A_{select}(i)} \exp(z_i \cdot z_k)}. \quad (6)$$

In general, ReCo injects the advantages of semantic relation from the hierarchical structure into the negative pairs sampling, and it can effectively regularize the overall contrastive learning:

$$\min_{f(\theta)} \mathcal{L}_{Supcon/PaCo} + \alpha \mathcal{L}_{ReCo}. \quad (7)$$

We note that ReCo is not sensitive to the weight α . We empirically set $\alpha = 1$.

5 Systematic Investigation on OmniBenchmark

We now report our thorough experimental studies on OmniBenchmark. These studies evaluate an extensive suite of recent representation learning methods. We first introduce these methods briefly (Sec. 5.1). Then, we quantify existing intuitions and reveal new insights from the results of the experiments (Sec. 5.2). Finally, we present the ablation studies of ReCo (Sec. 5.3), indicating that ReCo can boost the performance of the state-of-the-art supervised contrastive learning on both the ImageNet-1k and OmniBenchmark.

5.1 Models

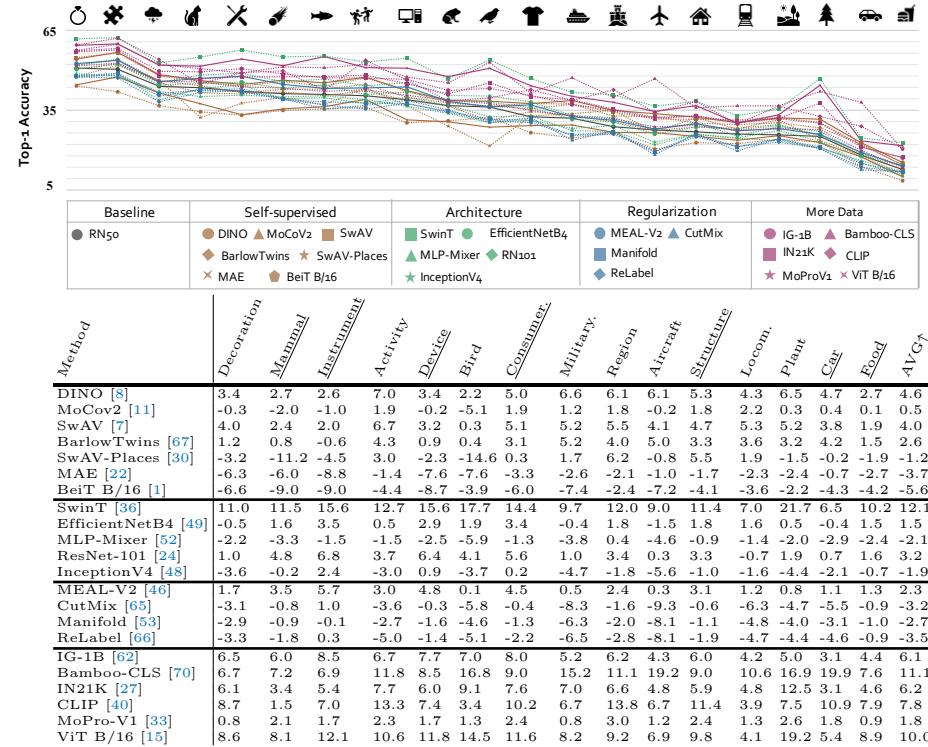
We benchmark 22 current models that are split into the following four categories referred to [44].

Self-supervised. We benchmark four types of self-supervised models, including contrastive (MoCov2 [11]), clustering-based (SwAV [7], DINO [8]), feature de-correlation (BarlowTwins [67]), masked autoencoder (MAE [22], BeiT [1]) models. These models are based on the ResNet-50 structure.

Architecture. We consider several architectures that include CNN based (ResNet-50 [24], ResNet-101 [24], EfficientNet-B4 [49] and Inception-v4 [48]), MLP based (MLP-Mixer [52]), Transformer based (Swin-T [36]). All these models are pre-trained on ImageNet-1k [43].

Regularization. ResNet-50 sized models with regularization techniques applied during the training phase include distillation (MEAL-V2 [46]), label augmentation (Manifold-MixUp [53], CutMix [65] and ReLabel [66]).

Table 1. Up: Linear classification Result On OmniBenchmark. We present the Top-1 linear probing accuracy on each realm dataset for 25 models listed in Sec. 5.1. **Down: Accuracy relative to the baseline ResNet50.** Accuracy relative to the baseline ResNet50 for all models, split across the four model categories. For the limited space, we only report the performance of models on 15 out of 21 realms. The complete results are shown in *Supplementary Material*. ImageNet-1k-seen-realms are marked in underline. Consumer. denotes consumer-goods, Locom. denotes locomotive.



Larger Scale Data. ResNet-50 Model pre-trained on the larger scale data (compared with ImageNet-1k (1M)) includes IG-1B [62] that is first pre-trained on IG-1B (1000 \times) and then fine-tuning on ImageNet-1k; CLIP that is image-text models pre-trained on WebImageText (400 \times); MoProV1 [33] that is pre-trained on the WebVision-V1 (2 \times); ViT-B/16 that is pre-trained on the ImageNet-22K (14 \times) [14] and then fine-tune on the ImageNet-1k; Bamboo-CLS [70] that is pre-trained on the Bamboo-CLS (65 \times).

5.2 Benchmarking Results

Table. 1 (up) reports top-1 accuracy of all models on different realms. Table. 1 (down) presents the performance of four model categories relative to the baseline ResNet50. Our main observations are as follows.

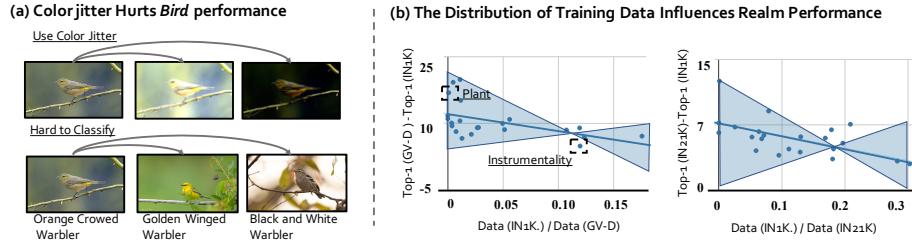


Fig. 4. The Highlighted Observation of Benchmark Results. (a) Though color jitters make models insensitive to color, it causes the model unable to classify among *Orange Crowed Warbler*, *Golden Winged Warbler* and *Black and White Warbler*. (b) On the *plant*, the amount of data in IN-1K is 0.08% of Bamboo-CLS; and on the *instrumentality*, the amount of data in IN-1K is 12.5% of Bamboo-CLS. Therefore, when it is compared with IN-1K model, the Bamboo-CLS model obtains larger top-1 accuracy gains on the *plant* than on the *instrumentality*. Each points denotes a realm.

The Selection of Pertaining Data Affects the Realm Performance. We hypothesize that using a pre-training dataset more similar than ImageNet-1k to a specific end realm dataset will produce a better encoder that performs better on that end realm dataset. We choose one model to test this hypothesis: SwAV-Places. Specifically, SwAV-Places is pre-trained on Places [71] for 200 epochs based on the SwAV framework, and we download the SwAV-Places from [30]. In Table. 1 (down) Self-supervised split, we find that for realms (*i.e.* structure and region) that are semantic similar to the Places dataset, SwAV-Places achieves the best performance compared with other Self-supervised learning (SSL) methods. This observation justifies our hypothesis.

Strong Augmentation of SSL Hurts Fine-Grained Realm Performance. SSL relies on strong data augmentation [7, 8, 11, 67] (*e.g.* color jitters) to learn visual representation, encouraging inductive bias. However, some data augmentation strategies may hurt the performance of fine-grain realm [60], *i.e.* Bird. As shown in Fig. 1 (down) Self-supervised split, except for masked encoder-based methods that use cropping-only augmentation, four out of five SSL methods achieve the worst performance on the bird realm. As illustrated in Fig. 4 (a), models that are insensitive to color transformation could struggle to classify between different warbler species.

Larger CNN Models Overfit to ImageNet-1k-seen-realms. As shown in Tab. 1 (down) Architecture split, we compare ResNet-50 with other CNN models with larger parameters. This figure illustrates that larger CNN models, *i.e.* Inception-v4 [48], EfficientNet-B4 [49] and ResNet-101 [24], achieve higher performance than ResNet-50 among most of realms. However, compared to the ResNet-50, we see larger gains that the these models exhibit on the realm: mammal, instrumentality, device, and consumer_goods are practically lost for the realm: aircraft, military_vehicle, and plant. We noted that mammal, instrumentality, device, and consumer_goods are the ImageNet-1k-seen-realm, but aircraft,

military_vehicle, and plant are ImageNet-1k-unseen-realm, we infer that larger CNN models might overfit more to ImageNet-1k-seen-realms.

Label Augmentation Techniques are Sensitive to Distribution Shift. Robustness to distribution shift is essential to evaluate model generalization ability [72]. Regularization methods exhibit strong performance gains over ResNet-50 on ImageNet-1k [53, 65, 66]. However, on the OmniBenchmark, these methods neither achieve good performance in the ImageNet-1k-unseen-realms nor show performance gain in the ImageNet-1k-seen-realms (*e.g.* instrumentality, consumer_goods, device), as shown in Table. 1 (down) Regularization split. We note that the ImageNet-1k and OmniBenchmark are built at different times, and the distribution shift thus exists in their data. Poor performance of label augmentation methods even on the ImageNet-1k-seen-realms of OmniBenchmark indicates that augmentation techniques are sensitive to the distribution shift. Unlike the larger parameters models that overfit to ImageNet-1k-seen-realm, which performs well on ImageNet-1k-seen-realm but lag in ImageNet-1k-unseen-realm, label augmentation techniques appear to overfit to ImageNet-1k instead of ImageNet-1k-seen-realm.

The Distribution of Training Data Influences Realm Performance. As shown in Fig. 4 (b), we argue that the distribution of training Data influences realm performance on OmniBenchmark. For example, ImageNet-1k has 1,250 times fewer data than Bamboo-CLS on the plant realm, when compared to ResNet-50, Bamboo-CLS obtains 16.9 points gain on that realm of OmniBenchmark. However, since ImageNet-1k has only 8 times fewer data than Bamboo-CLS on the instrumentality realm, Bamboo-CLS obtains only 6.9 points gain. Each point represents a specific realm.

OmniBenchmark is a Better Indicator for Generalization Ability. As shown in Bamboo-CLS paper [70], though DINO outperforms Bamboo-CLS on ImageNet-1k by 5.1%, Bamboo-CLS outperforms DINO by a large margin on other 9 out of 10 downstream tasks that include Food101 [5], SUN397 [59], OxfordPets [39] and *etc*, which indicate that the ImageNet-1k performance could not accurately indicates pre-trained models’ generalization ability in most downstream tasks. Surprisingly, Bamboo-CLS shows its better generalization ability than DINO on OmniBenchmark with 6.5 points gain. This observation align with the benchmarking results in Bamboo-CLS paper, and thus indicates that OmniBenchmark is a better indicator than ImageNet-1k for benchmarking pre-trained mode generalization ability.

5.3 Ablation Study of ReCo

We compare the performance of ReCo, and the state-of-the-art supervised contrastive learning method: parametric contrastive learning (PaCo) on both the pre-training dataset, *i.e.* ImageNet-1k, and OmniBenchmark. Better performance of ReCo indicates that it can effectively boost the current contrastive learning.

Implementation Details. Recently, parametric contrastive learning (PaCo) [13] achieves better ImageNet-1k performance than the conventional supervised contrastive learning (Supcon). To fair comparison, we reproduce PaCo and train

Table 2. The performance of SupCon, PaCo and ReCo on ImageNet-1k and OmniBenchmark. On the pre-training dataset, *i.e.* ImageNet-1k, both ReCo ResNet-50 and ResNet-101 models achieve 0.7 top-1 accuracy gain from the state-of-the-art supervised contrastive learning methods, *i.e.* PaCo. On the OmniBenchmark, ReCo ResNet-50 achieves 0.5 top-1 average performance gain from PaCo. Numbers in red are the performance gain on the same backbone network.

Method	Model	IN1K Top-1 ↑	IN1K Top-5 ↑	Omni. AVG↑
Supcon	ResNet-50	75.1	92.0	35.8
PaCo	ResNet-50	75.9	92.5	36.4
ReCo	ResNet-50	76.6 (+0.7)	93.0 (+0.5)	36.9 (+0.5)
Supcon	ResNet-101	78.9	94.4	37.9
PaCo	ResNet-101	79.1	94.4	38.4
ReCo	ResNet-101	79.8 (+0.7)	94.8 (+0.4)	38.7 (+0.3)

ReCo. Specifically, we use ResNet-50 [24] and ResNet-101 [24] as our backbones for experiments. The learning rate decays by a cosine scheduler from 0.1 to 0 with batch size 4096 on 32 GPUs in 200 epochs. These two models are trained using SGD optimizer with momentum $\mu = 0.9$.

Results on IN-1K and OmniBenchmark. The experimental results are summarized in Table 7. Our ResNet-50 model outperforms PaCo baseline models by 0.7%. And ReCo ResNet-50 improves the average top-1 performance of ReCo on the OmniBenchmark with a 0.5 point gain.

6 Conclusion

In our work, we develop a methodology for constructing a large-scale omni-realm benchmark, namely OmniBenchmark. Especially, as the basic building block of the OmniBenchmark, we define the visual realm indicated by expert knowledge, *i.e.* WordNet. Through extensive studies of recent representation learning methods, we find several insights. For example, the distribution of training data influences realm performance on our OmniBenchamrk. Besides, we propose a novel supervised contrastive learning, *i.e.* ReCo. ReCo selects the semantic dissimilar negative pairs rather than exhaustive sampling, boosting the performance of state-of-the-art supervised contrastive learning methods not only on ImageNet-1k but also on the OmniBenchmark. With the advent of the parameter-efficient tuning methods [25, 73, 74] in the vision task, we plan to evaluate various representation learning paradigm in the parameter-efficient tuning setting in the future. Overall, we hope our work could facilitate future research in omni-vision.

Acknowledgement

This work is supported by NTU NAP, MOE AcRF Tier 2 (T2EP20221-0033), and under the RIE2020 Industry Alignment Fund—Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). The corresponding author is Jing Shao.

References

1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers (2021) [10](#), [11](#)
2. Ben, H.: Duplicate image detection with perceptual hashing in python. <https://benhoyt.com/writings/duplicate-image-detection/#difference-hash-dhash> (2017) [6](#)
3. Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., Lord, N.A.: Making better mistakes: Leveraging class hierarchies with deep networks. In: CVPR. pp. 12506–12515 (2020) [4](#)
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021) [1](#)
5. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: ECCV (2014) [13](#)
6. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV. pp. 132–149 (2018) [4](#)
7. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020) [1](#), [2](#), [4](#), [10](#), [11](#), [12](#)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers (2021) [10](#), [11](#), [12](#)
9. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014) [1](#), [4](#)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607. PMLR (2020) [4](#)
11. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) [10](#), [11](#), [12](#)
12. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223. JMLR Workshop and Conference Proceedings (2011) [2](#)
13. Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J.: Parametric contrastive learning (2021) [3](#), [8](#), [13](#)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009) [3](#), [6](#), [11](#)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) [3](#), [4](#), [11](#)
16. Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with exemplar convolutional neural networks. TPAMI **38**(9), 1734–1747 (2015) [4](#)
17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2), 303–338 (2010) [6](#)
18. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPR workshop. pp. 178–178. IEEE (2004) [2](#)

19. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: ICCV. pp. 6391–6400 (2019) 1, 4, 7
20. Guo, J.N., Mao, X.L., Wei, W., Huang, H.: Intra-category aware hierarchical supervised document hashing. IEEE Transactions on Knowledge and Data Engineering (2022) 4
21. Guo, Y., Xu, M., Li, J., Ni, B., Zhu, X., Sun, Z., Xu, Y.: Hcsc: Hierarchical contrastive selective coding. arXiv preprint arXiv:2202.00455 (2022) 4, 9
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021) 10, 11
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020) 1, 2, 4
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 1, 3, 10, 11, 12, 14
25. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. arXiv preprint arXiv:2203.12119 (2022) 14
26. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020) 3, 8
27. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning (2020) 11
28. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. In: CVPR. pp. 1920–1929 (2019) 1
29. Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: CVPR. pp. 2661–2671 (2019) 1, 4
30. Kotar, K., Ilharco, G., Schmidt, L., Ehsani, K., Mottaghi, R.: Contrastive self-supervised representation learning models. arXiv preprint arXiv:2103.14005 (2021) 11, 12
31. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 2
32. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: ECCV. pp. 577–593. Springer (2016) 4
33. Li, J., Xiong, C., Hoi, S.C.: Mopro: Webly supervised learning with momentum prototypes. ICLR (2021) 11
34. Li, J., Zhou, P., Xiong, C., Socher, R., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. arXiv preprint arXiv:2005.04966 (2020) 4
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014) 6
36. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021) 10, 11
37. Miller, G.A.: WordNet: An electronic lexical database. MIT press (1998) 2
38. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: CVPR. pp. 6707–6717 (2020) 4
39. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR. pp. 3498–3505. IEEE (2012) 13
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 1, 2, 4, 11

41. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters (2017) [4](#)
42. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International Conference on Machine Learning. pp. 5389–5400. PMLR (2019) [1](#), [4](#)
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015) [2](#), [5](#), [10](#)
44. Sarıyıldız, M.B., Kalantidis, Y., Larlus, D., Alahari, K.: Concept generalization in visual representation learning (2021) [4](#), [7](#), [10](#)
45. Shao, J., Chen, S., Li, Y., Wang, K., Yin, Z., He, Y., Teng, J., Sun, Q., Gao, M., Liu, J., et al.: Intern: A new learning paradigm towards general vision. arXiv preprint arXiv:2111.08687 (2021) [6](#)
46. Shen, Z., Savvides, M.: Meal v2: Boosting vanilla resnet-50 to 80accuracy on imagenet without tricks (2021) [10](#), [11](#)
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) [1](#), [4](#)
48. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning (2016) [10](#), [11](#), [12](#)
49. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks (2020) [10](#), [11](#), [12](#)
50. Tanon, T.P., Weikum, G., Suchanek, F.M.: Yago 4: A reason-able knowledge base. The Semantic Web **12**¹²**3**, 583 – 596 (2020) [5](#)
51. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? (2020) [8](#)
52. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A.: Mlp-mixer: An all-mlp architecture for vision (2021) [10](#), [11](#)
53. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Ben gio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: ICML (2019) [10](#), [11](#), [13](#)
54. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014) [2](#)
55. Wang, W., Zhou, T., Qi, S., Shen, J., Zhu, S.C.: Hierarchical human semantic parsing with comprehensive part-relation modeling. TPAMI (2021) [4](#)
56. Wang, Y., Hu, Q., Zhu, P., Li, L., Lu, B., Garibaldi, J.M., Li, X.: Deep fuzzy tree for large-scale hierarchical visual classification. IEEE Transactions on Fuzzy Systems **28**(7), 1395–1406 (2019) [4](#)
57. Wang, Y., Wang, Z., Hu, Q., Zhou, Y., Su, H.: Hierarchical semantic risk minimization for large-scale classification. IEEE Transactions on Cybernetics (2021) [4](#)
58. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR. pp. 3733–3742 (2018) [4](#)
59. Xiao, J., Ehinger, K.A., Hays, J., Torralba, A., Oliva, A.: Sun database: Exploring a large collection of scene categories. IJCV **119**(1), 3–22 (2016) [13](#)
60. Xiao, T., Wang, X., Efros, A.A., Darrell, T.: What should not be contrastive in contrastive learning (2021) [12](#)
61. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR. pp. 10687–10698 (2020) [4](#)

62. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019) [1](#), [3](#), [4](#), [11](#)
63. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 547–558 (2020) [5](#)
64. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: CVPR. pp. 5525–5533 (2016) [5](#)
65. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features (2019) [10](#), [11](#), [13](#)
66. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: from single to multi-labels, from global to localized labels. arXiv preprint arXiv:2101.05022 (2021) [1](#), [10](#), [11](#), [13](#)
67. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction (2021) [10](#), [11](#), [12](#)
68. Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A.S., Neumann, M., Dosovitskiy, A., et al.: A large-scale study of representation learning with the visual task adaptation benchmark. arXiv preprint arXiv:1910.04867 (2019) [1](#), [2](#), [4](#)
69. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV. pp. 649–666. Springer (2016) [4](#)
70. Zhang, Y., Sun, Q., Zhou, Y., He, Z., Yin, Z., Wang, K., Sheng, L., Qiao, Y., Shao, J., Liu, Z.: Bamboo: Building mega-scale vision dataset continually with human-machine synergy. arXiv preprint arXiv:2203.07845 (2022) [6](#), [11](#), [13](#)
71. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. TPAMI (2017) [6](#), [12](#)
72. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey (2021) [13](#)
73. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134 (2021) [14](#)
74. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) [14](#)
75. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.V.: Rethinking pre-training and self-training. arXiv preprint arXiv:2006.06882 (2020) [4](#)