MEASURING THE INTERPRETABILITY OF UNSUPERVISED REPRESENTATIONS VIA QUANTIZED REVERSE PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised visual representation learning has attracted significant research interest. While the most common way to evaluate self-supervised representations is through transfer to various downstream tasks, we instead investigate the problem of measuring their interpretability, *i.e.* understanding the semantics encoded in the raw representations. We formulate the latter as estimating the mutual information between the representation and a space of manually labelled concepts. To quantify this we introduce a decoding bottleneck: information must be captured by simple predictors, mapping concepts to clusters of data formed in representation space. This approach, which we call *reverse linear probing*, provides a single number sensitive to the semanticity of the representation. This measure is also able to detect when the representation correlates with combinations of labelled concepts (*e.g.* "red apple") instead of just individual attributes ("red" and "apple" separately). Finally, we also suggest that supervised classifiers can be used to automatically label large datasets with a rich space of attributes. We use these insights to evaluate a large number of self-supervised representations, ranking them by interpretability, and highlight the differences that emerge compared to the standard evaluation with linear probes.

1 INTRODUCTION

Relying on black-box models such as deep networks comes sometimes with significant methodological and ethical challenges. This is particularly true for *unsupervised* and *self-supervised* models which are learned without human supervision. While these models perform increasingly well in downstream applications, often outperforming supervised counterparts, there is very little understanding of what they learn, making their real-life deployment risky.

In this paper, we thus consider the problem of characterizing the *meaning* of data representations, with a particular focus on unsupervised and self-supervised representations of images. Given a representation f mapping images x to representation vectors $f(x) \in \mathbb{R}^D$, our goal is to find whether these vectors contain human-interpretable information. This is usually done by finding a relationship or correspondence between $f(x)$ and human-provided descriptions $y(x)$ of the images, essentially translating the information from the representation space to a concept space. Consider for example the popular linear probing method (Alain and Bengio, 2017). Given a large dataset $\hat{\mathcal{X}}$ of images with corresponding manual annotations $y(x)$, one learns linear classifiers (*probes*) to map the feature vectors $f(x)$ to labels $y(x)$, measuring the resulting classification accuracy. If the predictor is accurate, then one can argue that the representation captures the corresponding concept y.

A limitation of this type of approaches is that one can only discover in the representation meanings that are represented in the available data annotations. In order to maximize the semantic coverage of the analysis, it is thus customary to combine annotations for several type of attributes, such as the object class, colour, size, etc. In this case, y is a vector of image attributes, and one can compute several linear probes to predict each individual attribute y_i in turn. By doing so, however, attributes are treated independently, which is unsatisfactory. In order to obtain a single number assessing the overall semanticity of a representation, one must combine the prediction accuracies of several independent classifiers, and there is no natural way of doing so (*e.g.* a simple average

would not account for the different complexity of the different prediction tasks). Furthermore, the representation might be predictive of *combinations* of attributes, *i.e*. it might understand the concept of "red apple" without necessarily understanding the individual concepts, *e.g*. "red" and "apple". While it is in principle possible to test any combination of attributes via linear probing, most attribute combinations are too rare to generate significant statistics for this analysis.

In this paper, we thus propose a complementary assessment strategy that addresses these shortcomings. Compared to linear probes, we start by considering the reverse prediction problem, mapping label vectors $y(x)$ to representation vectors $f(x)$. A key advantage is that the entire attribute vector is used to perform this prediction, thus accounting for attribute combinations naturally.

Next, we consider the challenge of deriving a quantity that allows to compare representations based on the performance of these reverse predictors. Obviously a simple metric such as the average L^2 prediction error is meaningless as its magnitude would be affected by irrelevant factors such as the scale of the representation vectors. To solve this problem in a principled manner, we propose to consider instead the *mutual information* $I(f_K(x), y(x))$ between the label vectors and a quantized version of the representation vectors. This approach, which we justify from the viewpoint of information theory, essentially measures whether the representation groups the data in a human-interpretable manner.

We use our approach to evaluate a large number of self-supervised and supervised image representations, which results in different ranking of methods compared to standard linear probing. We also show that these representations find many interpretable concepts that extend beyond the classes of ImageNet. Interestingly, methods that perform better on standard ImageNet linear probing, seem to mainly learn better "ImageNet concepts" at the cost of diminished performance on other, more general concepts. We investigate these differences quantitatively and qualitatively, which further support the use of our method to understand representations.

2 Related Work

Self-supervised representation learning (SSL) In this paper, we focus on self-supervised methods that learn representations from image data. Early self-supervised learning approaches devise numerous pretext tasks — such as colorization, predicting image rotations or solving image puzzles — to learn useful representations from unlabeled images (Gidaris et al., 2018; Pathak et al., 2016; Noroozi and Favaro, 2016; Doersch et al., 2015; Larsson et al., 2016; Zhang et al., 2017; 2016). More recent approaches follow the contrastive learning paradigm (Chen et al., 2020a; Frankle et al., 2020; Kalantidis et al., 2020; He et al., 2020; Chen et al., 2020b; Hénaff et al., 2019; Oord et al., 2018; Misra and Maaten, 2020; Tian et al., 2019; 2020; Wu et al., 2018), where the goal is to discriminate instances belonging to the same image from negative samples. Other methods (Zbontar et al., 2021; Chen and He, 2020; Grill et al., 2020; Caron et al., 2021) eliminate the need for negative samples. While there exists some theoretical approaches for explaining why the constrastive learning works (Arora et al., 2019; Wang and Isola, 2020; Tschannen et al., 2020; Tsai et al., 2020), in most cases the focus lies on showing empirical improvement in downstream tasks.

Clustering A related part of literature deals with unsupervised image clustering. Early approaches include using autoencoders Hinton and Salakhutdinov (2006), agglomerative clustering (Bautista et al., 2016) and partially ordered sets of hand-crafted features (Bautista et al., 2017). More recent methods combine representation learning with clustering, using mutual information (Ji et al., 2019; Hu et al., 2017), K-means (Caron et al., 2018; Zhan et al., 2020), prototypes (Li et al., 2021) or optimal transport (Asano et al., 2020; Caron et al., 2020). There are also methods which build on strong feature representations and, at a second stage, either only cluster (Yan et al., 2020a) or cluster and further train and refine the network (Van Gansbeke et al., 2020; Lee et al., 2019). Besides evaluating the representations, the pseudolabels such methods produce can be used for evaluating representations by measuring the correlation to the ground-truth labels.

Evaluation and analysis of SSL There are two main directions in evaluating self-supervised learning methods. The first one is using the pre-trained representations as an initialization for subsequent supervised training, *e.g*. for object detection. The second and more popular approach is to train a linear classifier on the embedding spaces which are kept frozen after pre-training. As a result, several benchmarking suites with various datasets and tasks have been proposed (Zhai et al., 2019;

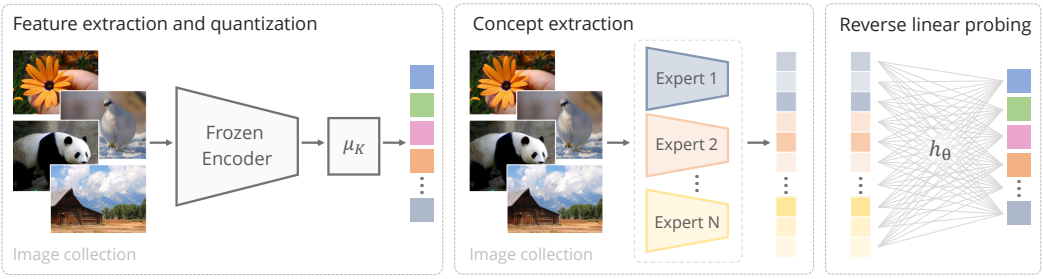


Figure 1: Overview of our approach. (1) we evaluate pre-trained SSL models on an image collection, extracting and quantizing feature vectors to obtain clusters ( ), (2) we label the image data with a diverse set of concepts ( ) from expert models trained with supervision on external data sources, and (3) we train a linear model h_θ to map concepts to clusters, measuring the mutual information between the representation and human-interpretable concepts.

Goyal et al., 2019b; Van Horn et al., 2021; Kotar et al., 2021). Complementary to this, clustering data in the frozen feature space has also been proposed as another evaluation metric (Zheltonozhskii et al., 2020; Sariyildiz et al., 2020). In addition, downstream dataset (Ericsson et al., 2021) and pretraining dataset (Zhao et al., 2021; Cole et al., 2021) dependencies have been evaluated. Finally, another line of work has been recently investigating self-supervised feature spaces to understand separability (Sehwag et al., 2020), concept generalization (Sariyildiz et al., 2020) and the effect of balanced vs. long-tailed training data (Kang et al., 2020).

Interpretability Although a large amount of work has studied feature representations in Convolutional Neural Networks (CNNs), it is heavily focused on models trained with full supervision. Zeiler and Fergus (2014); Zhou et al. (2014) analyze networks by visualizing the most activating patches, while activation maximization methods Mahendran and Vedaldi (2016a;b); Nguyen et al. (2017; 2016); Olah et al. (2017); Simonyan et al. (2014) generate inputs to most strongly activate specific neurons. More recently, Kim et al. (2018) relate human-interpretable concepts to labelled concepts, Ghorbani et al. (2019) propose a method to automatically assign concepts to image segments and Yeh et al. (2020) analyze the completeness of these concepts for explaining the CNN.

Closer to our method, Bau et al. (2017) propose to quantify the interpretability of individual units in deep networks by measuring the overlap between each unit and a set of densely annotated concepts, while Alain and Bengio (2017) probe entire representations at intermediate layers using linear classifiers. Similarly, Zhou et al. (2018) propose to decompose feature vectors into a set of elementary and interpretable components and show decomposed Grad-CAM heat maps (Selvaraju et al., 2017) for concepts that significantly contribute to the prediction. With the exception of (Laina et al., 2020; Bau et al., 2017; Fong and Vedaldi, 2018), relatively little work has been done to understand the emergence of interpretable visual concepts in self-supervised representations specifically. In particular, Laina et al. (2020) quantify the learnability and describability of unsupervised image groupings, by measuring how successful humans are at understanding the learned concept from a set of provided examples. Instead, our approach does not depend on human input and is thus easily scalable to a wide range of methods and number of classes.

3 METHOD

We are interested in measuring the semanticity of data representations. A representation f is a map $f : \mathcal{X} \to \mathbb{R}^D$ that encodes data samples $x \in \mathcal{X}$ as vectors $f(x) \in \mathbb{R}^D$. In this paper, we assume that the data are images $\mathcal{X} = \mathbb{R}^{3\times H\times W}$ and f is a deep neural network, but other choices are possible.

The "meaning" or semantic content of an image x can be characterized by obtaining a label or description $y(x)$ of the image from a human annotator. If the representation $f(x)$ captures the meaning of the image too, then it should be predictive of the description $y(x)$. The mutual information $I(f(x), y(x))$ provides a natural measure of predictivity, and it is thus tempting to use this quantity as a measure of the overall semanticity of the representation. However, information as such is not a useful measure in our context. To see why, note that, due to the data processing inequality

($I(f(x), y(x)) \leq I(x, y(x))$), the information is maximized by observing the raw image, *i.e.* by the identity representation $f(x) = x$. Since processing the data cannot increase its information content, useful representation must preserve information while *also* making it easier to decode and act on it.

There are many possible definitions of what constitutes 'simple decoding' — some of which will be further discussed later in this section. Here we propose a definition that stems from the interpretation of differential entropy as the limit of discrete entropy for quantized versions of the corresponding variables (Cover and Thomas, 2006). In our case, the description $y(x)$ is already discrete, but the representation vectors are continuous. We thus propose to vector-quantize the representation and to use the resulting discrete information as a measure of semanticity.

Intuitively, in the limit of infinitely-fine quantization, this quantity reduces (up to a normalizing term) to the mutual information above. For a finite number of clusters, however, a large value of information means that the representation groups images in a way that makes sense to a human observer. In this case, the 'simple decoding' amounts to grouping images based on the L^2 distance between their representation vectors.

Formally, we use a vector quantization algorithm such as K-means (Lloyd, 1982) to learn a quantizer function $\mu_K : \mathbb{R}^D \to \{1, \dots, K\}$ for the representation vectors and estimate the mutual information $I(\mu_K(f(x)), y(x))$ by rewriting it as:

$$I(f_K(x), y(x)) = H(f_K(x)) - H(f_K(x) \mid y(x)), \quad f_K(x) = \mu_K(f(x)). \tag{1}$$

In the above equation, the first term denotes the entropy of the cluster assignments. In practice, given a sample dataset $\hat{\mathcal{X}}$ of images, we first run K-means to compute the quantizer μ_K, and then compute the frequency of cluster assignments $f_K(x)$, $x \in \mathcal{X}$ to calculate the entropy. The second term in Eq. (1) is the conditional entropy:

$$H(f_K(x) \mid y(x)) = \mathbb{E}_p[-\ln p(f_K(x) \mid y(x))] \leq \mathbb{E}_p[-\ln q(f_K(x) \mid y(x))]$$

where p is the true joint distribution between variables $f_K(x)$ and $y(x)$. While this is difficult to compute, we can upper-bound it by considering an auxiliary posterior distribution q interpreting the conditional entropy as the cross-entropy loss of a *predictor* h_θ, parametrized by θ, that maps labels $y(x)$ to clusters $f_K(x)$. We minimize the gap by learning this predictor in the usual way.

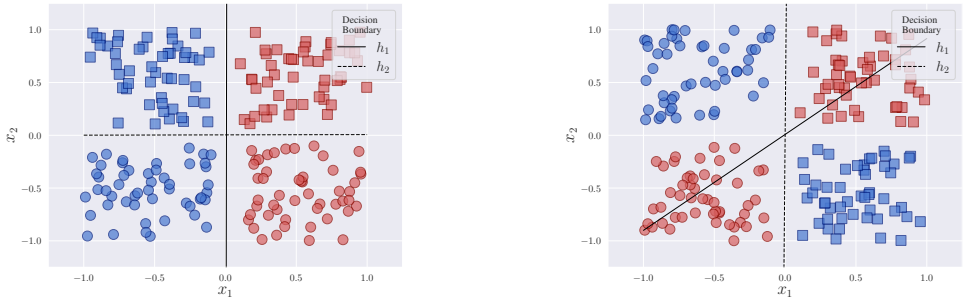
In practice, learning the predictor may result in overfitting, which would lead to an over-optimistic estimate of the mutual information. We address this issue by learning the predictor h_θ on a subset $\hat{\mathcal{X}}' \subset \hat{\mathcal{X}}$ of the data and evaluating its cross-entropy loss on the remaining subset $\hat{\mathcal{X}} - \hat{\mathcal{X}}'$. Importantly, we consider a linear predictor (*probe*) for h_θ, for speed and to further reduce the risk of overfitting. We refer to this as quantized reverse probing, as it maps semantic concepts to quantized data representations.

Obtaining labelled data via automatic experts Our method, similar to prior works, requires a dataset $x \in \hat{\mathcal{X}}$ of images equipped with manually-provided labels $y(x) \in \mathcal{Y}$. Because we do not know *a-priori* which concepts may be capture by a given representation $f(x)$ under analysis, these labels should provide a good coverage for a number of different concepts (e.g., textures, object types, parts, etc.). A good example of such a dataset is Broden (Bau et al., 2017).

We can further and cheaply increase the coverage of the label space by *predicting* the labels $y(x)$ automatically via a battery of expert classifiers learned using full manual supervision. The noise of these automated predictions is small compared to the noise in the correlation between the concepts and the unsupervised/self-supervised visual representations under study.

Number of clusters The number of clusters K controls the size of the bottleneck. Mutual information increases monotonically as K increases, as we show in the Appendix (Fig. 6); when comparing representations, it is thus important to fix a value of K and use it consistently. For added convenience, in practice we report a normalised value of information (NMI).

Relation to linear probing Linear probing is a common approach for assessing representations, related to our method. A linear probe is a simple predictor $h_i \circ f(x) \approx y_i(x)$ that maps the representation $f(x)$ to a specific concept y_i (e.g., a binary attribute or a class label). The idea is that the representation captures concept y_i if the probe performs well. Similar to our approach, the simplicity of the probes (imposed via linearity and sometimes sparsity) is a necessary bottleneck for interpretability: without such a constrain, the best possible probe would take the raw image as input.



(a) All attributes are linearly separable (Acc[h_1] = Acc[h_2] = 100%).

(b) Color attribute not linearly separable (Acc[h_1] = 50%, Acc[h_2] = 100%).

Figure 2: **Forward vs. reverse probing.** A dataset $\hat{\mathcal{X}}$ embedded in $\mathbb{R}^2$ by two different representations. Each data point has two binary attributes, **color** $y_1(x)$: red or blue and **shape** $y_2(x)$: □ or ○. Both representations separate these attributes; however, color is not linearly separable in (b), so forward linear probing cannot recover this relationship. On the contrary, reverse probing easily discovers that all *combinations* of shape and color map to well separated clusters.

Table 1: Different types of concepts and related datasets.

Categories	**Tasks/Datasets**
OBJECTS	IN-1k (Russakovsky et al., 2015); Open Images (Kuznetsova et al., 2020); MSCOCO (things) (Lin et al., 2014)
SCENE	Places-365 (Zhou et al., 2016); Scene attributes (Patterson and Hays, 2012)
MATERIAL	MINC (Bell et al., 2015); MSCOCO (stuff) (Caesar et al., 2018b)
TEXTURE	DTD (Cimpoi et al., 2014); Color
OTHER	Text detection; Sentiment; Photographic style

Our method is complementary to linear probes, with some advantages. First, it provides a single principled figure characterising the entire representation instead of a separate figure for each predicted attribute y_i. Second, by learning predictors in the label-to-representation direction $\mathcal{Y} \to \mathcal{X}$, it handles naturally combinations of multiple attributes, whereas linear probes, by operating in the $\mathcal{X} \to \mathcal{Y}$ direction, amount to a collection of independent attribute-specific predictors. In this manner, our approach can efficiently correlate the representation with *combinations of attributes*, for instance testing for "red circle" instead of, separately, for "red" and "circle", as in the example of Fig. 2. Third, also owing to this efficiency, our measure can be computed much faster than linear probes, sometimes cutting down the time required from hours to minutes.

4 EXPERIMENTS

We use our method to evaluate a wide range of recent self-supervised representation learning and clustering techniques. In the following, we compare these techniques according to the proposed criterion, which results in a new ranking for the state-of-the-art approaches. Further, we show qualitatively how our method discovers elementary concepts which are encoded in the raw representations.

4.1 IMPLEMENTATION DETAILS

Attributes and expert models As a first step in our approach, we collect a set of attributes which we expect to find encoded in the self-supervised representations. As shown in Table 1, these include semantic object categories, scene types and attributes, materials and textures and possibly other information about the photographic style or overall sentiment of an image. For each attribute category, we look for relevant human-annotated datasets and expert models trained on these datasets as a proxy to human knowledge. This allows us to then extract information related to such attributes on a different dataset or images in the wild; in our experiments we focus on ImageNet 2012 split (IN-1k) (Deng et al., 2009) as the target data due to the large availability of pre-trained self-supervised

Table 2: Evaluation of self-supervised learning methods on ImageNet-1k (Russakovsky et al., 2015). Lin. Acc. refers to the linear classification accuracy reported by the respective source for each model. The remaining metrics are computed for our approach and reported as mean $\pm$ std over 5 clusterings.

	Model		Lin. Acc.	NMI	AMI	Top-1	mAP
			ResNet-50 ($1\times$)				
Epoch 200	Jigsaw	(Goyal et al., 2019a)	46.58	37.36 $\pm$ 0.04	22.17 $\pm$ 0.04	9.80 $\pm$ 0.04	5.95 $\pm$ 0.04
	ClusterFit	(Yan et al., 2020b)	53.63	51.10 $\pm$ 0.07	38.78 $\pm$ 0.10	30.06 $\pm$ 0.09	25.41 $\pm$ 0.11
	CMC	(Tian et al., 2019)	58.60	51.96 $\pm$ 0.06	39.33 $\pm$ 0.07	29.01 $\pm$ 0.15	25.77 $\pm$ 0.23
	MoCo-v1	(He et al., 2020)	60.60	55.94 $\pm$ 0.05	44.39 $\pm$ 0.07	34.36 $\pm$ 0.13	33.36 $\pm$ 0.17
	SeLa-v1	(Asano et al., 2020)	61.50	59.91 $\pm$ 0.04	49.02 $\pm$ 0.08	41.22 $\pm$ 0.13	39.94 $\pm$ 0.30
	SimCLR	(Chen et al., 2020a)	66.61	63.73 $\pm$ 0.07	54.87 $\pm$ 0.09	47.93 $\pm$ 0.09	53.95 $\pm$ 0.08
	MoCo-v2	(Chen et al., 2020b)	67.70	64.76 $\pm$ 0.05	56.05 $\pm$ 0.05	47.53 $\pm$ 0.10	51.97 $\pm$ 0.19
	InfoMin	(Tian et al., 2020)	70.10	65.16 $\pm$ 0.07	57.27 $\pm$ 0.05	48.11 $\pm$ 0.11	54.79 $\pm$ 0.13
	MoCHi	(Kalantidis et al., 2020)	67.60	65.27 $\pm$ 0.09	56.94 $\pm$ 0.13	49.26 $\pm$ 0.22	56.15 $\pm$ 0.40
	PCL-v2	(Li et al., 2021)	67.60	68.89 $\pm$ 0.06	62.13 $\pm$ 0.09	53.72 $\pm$ 0.08	61.26 $\pm$ 0.16
	SwAV	(Caron et al., 2020)	73.90	69.18 $\pm$ 0.05	61.68 $\pm$ 0.06	55.85 $\pm$ 0.09	62.77 $\pm$ 0.27
	OBoW	(Gidaris et al., 2021)	73.80	**71.50** $\pm$ **0.07**	64.09 $\pm$ 0.09	57.25 $\pm$ 0.13	61.67 $\pm$ 0.22
Epoch 400	SimCLR	(Chen et al., 2020a)	67.71	64.88 $\pm$ 0.09	56.45 $\pm$ 0.12	50.03 $\pm$ 0.17	56.92 $\pm$ 0.21
	SwAV	(Caron et al., 2020)	74.60	69.09 $\pm$ 0.12	61.66 $\pm$ 0.18	56.06 $\pm$ 0.22	62.78 $\pm$ 0.25
	SeLa-v2	(Asano et al., 2020)	71.80	70.04 $\pm$ 0.06	63.43 $\pm$ 0.07	58.47 $\pm$ 0.15	**68.33** $\pm$ **0.34**
	DeepCluster-v2	(Caron et al., 2020)	74.32	70.85 $\pm$ 0.07	64.01 $\pm$ 0.12	58.91 $\pm$ 0.15	67.60 $\pm$ 0.27
Epoch 800	SimCLR	(Chen et al., 2020a)	69.68	65.63 $\pm$ 0.14	57.48 $\pm$ 0.15	51.36 $\pm$ 0.17	59.11 $\pm$ 0.31
	PIRL	(Misra and van der Maaten, 2020)	69.90	65.92 $\pm$ 0.05	58.79 $\pm$ 0.09	51.95 $\pm$ 0.07	60.75 $\pm$ 0.11
	DINO	(Caron et al., 2021)	75.30	68.73 $\pm$ 0.08	61.18 $\pm$ 0.11	55.33 $\pm$ 0.21	59.87 $\pm$ 0.21
	SwAV	(Caron et al., 2020)	75.30	68.79 $\pm$ 0.05	61.19 $\pm$ 0.08	55.73 $\pm$ 0.06	62.17 $\pm$ 0.28
	MoCHi	(Kalantidis et al., 2020)	69.20	69.00 $\pm$ 0.07	61.77 $\pm$ 0.06	55.23 $\pm$ 0.07	63.81 $\pm$ 0.09
	MoCo-v2	(Chen et al., 2020b)	71.10	69.02 $\pm$ 0.07	61.55 $\pm$ 0.11	54.17 $\pm$ 0.17	60.44 $\pm$ 0.10
	InfoMin	(Tian et al., 2020)	73.00	69.20 $\pm$ 0.02	62.54 $\pm$ 0.04	55.15 $\pm$ 0.10	63.84 $\pm$ 0.10
	DeepCluster-v2	(Caron et al., 2020)	75.18	69.44 $\pm$ 0.04	62.12 $\pm$ 0.05	57.01 $\pm$ 0.08	63.97 $\pm$ 0.26
Epoch 1k	Barlow Twins	(Zbontar et al., 2021)	73.50	69.37 $\pm$ 0.08	61.69 $\pm$ 0.13	56.84 $\pm$ 0.18	61.32 $\pm$ 0.23
	MoCHi	(Kalantidis et al., 2020)	70.60	70.16 $\pm$ 0.06	63.37 $\pm$ 0.09	57.08 $\pm$ 0.16	66.18 $\pm$ 0.11
	BYOL	(Grill et al., 2020)	74.40	70.48 $\pm$ 0.07	63.12 $\pm$ 0.10	58.36 $\pm$ 0.09	63.26 $\pm$ 0.25
	MoCo-v3	(Chen et al., 2021)	74.60	**71.45** $\pm$ **0.06**	**64.49** $\pm$ **0.09**	**59.58** $\pm$ **0.11**	64.95 $\pm$ 0.43
	Supervised	(He et al., 2016)	-	83.20 $\pm$ 0.06	78.82 $\pm$ 0.07	76.16 $\pm$ 0.14	78.53 $\pm$ 0.15
			ViT-Base/16				
	MoCo-v3	(Chen et al., 2021)	76.70	79.06 $\pm$ 0.04	73.67 $\pm$ 0.05	70.51 $\pm$ 0.09	74.39 $\pm$ 0.28
	DINO	(Caron et al., 2021)	78.20	81.46 $\pm$ 0.08	76.70 $\pm$ 0.11	72.95 $\pm$ 0.14	76.44 $\pm$ 0.18
	Supervised	(Touvron et al., 2021)	-	94.36 $\pm$ 0.11	93.13 $\pm$ 0.14	92.02 $\pm$ 0.19	80.82 $\pm$ 0.15

models on this dataset. In all cases, we concatenate all available attributes and form M-dimensional binary vectors denoting the presence or absence of each attribute class in the image. We provide details on the specific models and datasets used in Appendix A.

Since we primarily focus on ImageNet, it is natural to also use the existing 1000 human-annotated labels. As IN-1k already includes several fine-grained categories, we have not included further experts from common fine-grained recognition datasets, such as iNaturalist (Van Horn et al., 2018).

Linear model For each method, we train a linear model mapping a set of input attributes to clusters, which are obtained by K-means clustering on top of fixed representations. We train K-means for 100 steps and choose the best out of 5 runs. The linear model is trained for 100 epochs with SGD on a trainval split created by splitting the cluster assignments with a 80/20 ratio. Further implementation details are provided in Appendix A.

4.2 Comparison of self-supervised representations

We use reverse probing to evaluate recent self-supervised techniques in Table 2 and rank them based on their semanticity. We obtain Jigsaw, SimCLR, ClusterFit from the VISSL model zoo (Goyal et al., 2021) and for the rest we use the official implementations. For each method, we freeze the pre-trained model and extract feature vectors on the IN-1k data. We then perform K-means on the

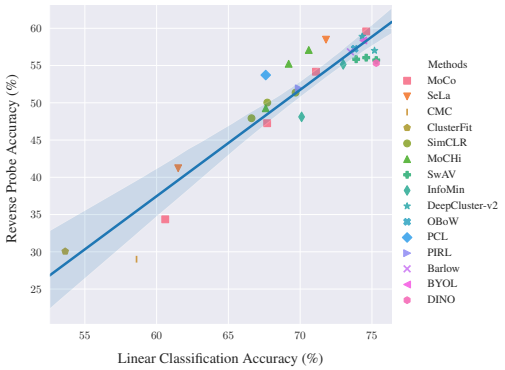


Figure 3: Linear classification accuracy on IN-1k vs. classification accuracy of our probe (top-1). A linear regression model is fit to the data, suggesting strong correlation. Each point corresponds to a ResNet-50-based model in Table 2.

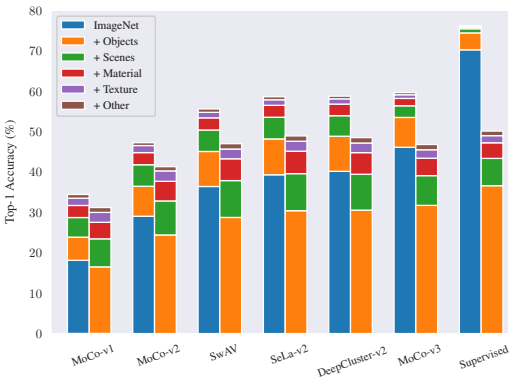


Figure 4: Contribution of each expert group to the overall accuracy for selected methods. For each bar a linear model is trained using as input the current expert group and all previous ones (*e.g.* green bar: IN-1K+OBJECTS+ SCENES).

features (where $K = 1000$) to obtain cluster assignments and train our linear model (h_θ) using *all* categories from Table 1 as input and the cluster assignments as targets. We repeat this $5\times$ to obtain different clusterings, each used to train a linear model from scratch and compute the mean and standard deviation over the 5 clusterings for each method. In addition to a normalized measure of mutual information (NMI), we also report the adjusted mutual information (AMI), classification accuracy (top-1) and mean average precision (mAP). We also note the linear classification accuracy, as reported by the respective source for each method. For fairness of comparisons, we evaluate ResNet-50 models and ViT-Base models and report the number of training epochs for each model. For reference, we also evaluate the supervised counterparts in the same way as for the self-supervised methods, *i.e.* by clustering features to obtain the target pseudo-classes.

Overall, we find that there is generally a strong correlation between our approach and the common linear evaluation protocol (Fig. 3). However, we make the following additional observations:

Ranking Despite the strong correlation to linear classification accuracy, we obtain a different ranking following our approach, *i.e.* from the perspective of representational interpretability. In particular, state-of-the-art methods DINO, SwAV and BYOL fall slightly behind in our evaluation. The most recent version of MoCo (v3) performs the best across three metrics, with DeepCluster-v2 and OBoW ranking in the second place (across ResNet-50-based models).

Number of pre-training epochs For some methods, learned weights are available at various epochs. In terms of interpretability, we observe no benefit in longer pre-training in the case of DeepCluster-v2 and SwAV, despite the solid performance gains on standard benchmarks. However, there is a noticeable gain for MoCo-v2, MoCHi and SimCLR between 200 and 800 epochs. Notably, one of the best performing methods, OBoW is only trained for 200 epochs.

Clustering-based vs contrastive approaches Methods such as SeLa, PCL and DeepCluster-v2, *i.e.* those that use a clustering mechanism during training, have generally more interpretable representations, that is they rank higher than methods with similar linear classification accuracy, lying above the line in Fig. 3.

Architecture Self-supervised methods that use a ViT (Touvron et al., 2021; Dosovitskiy et al., 2021) backbone show a significant boost in all metrics, even more so than linear classification accuracy suggests. However, a fair comparison between MoCo-v3 and DINO is not possible due to pre-training for a different number of epochs (300 vs. 800).

4.3 EXPERT BREAKDOWN

Our next goal is to understand the effect of different concepts on explaining the representation, *i.e.* we want to answer the question: to which degree does the representation *know* certain concepts,

Table 3: Change in NMI when learning the reverse probe with individual concept groups on top of the human-annotated IN-1K labels.

Method		IN-1K	OBJECTS	SCENE	MATERIAL	TEXTURE	OTHER
MoCo-v1	(He et al., 2020)	47.70	+2.73	+2.92	+2.93	+1.04	+0.21
MoCo-v2	(Chen et al., 2020b)	57.39	+2.43	+2.09	+2.03	+0.43	-0.24
SwAV	(Caron et al., 2020)	61.94	+2.42	+1.92	+2.01	+0.23	-0.20
SeLa-v2	(Asano et al., 2020)	63.29	+2.23	+1.69	+1.69	+0.03	-0.45
DeepCluster-v2	(Caron et al., 2020)	64.29	+2.17	+1.54	+1.70	+0.12	-0.35
OBoW	(Gidaris et al., 2021)	64.73	+2.77	+1.82	+2.13	+0.11	-0.15
MoCo-v3	(Chen et al., 2021)	67.69	+1.51	+0.32	+0.21	-0.36	-0.38

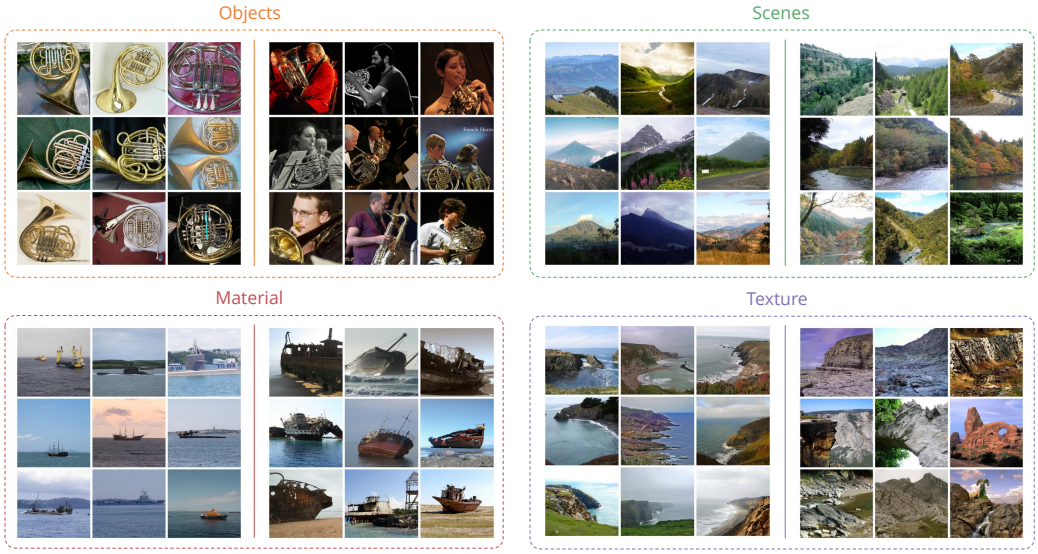


Figure 5: Qualitative examples. We show pairs of clusters for which the estimated confusion was significantly reduced after training the linear model with the corresponding concepts. We refer to Appendix C for more examples.

e.g. material? As we are focusing on self-supervised representations here, the answer to this question is far from obvious. We again measure this via the predictive power of our probe, which we now train individually for each group of concepts. These vary in nature; from highly semantic object categories (*e.g.* `dog`, `avocado`) to lower-level features such as material (*e.g.* `wood`, `brick`), texture (*e.g.* `bubbly`, `striped`) and color (*e.g.* `red`).

In Fig. 4, we show how the different concept groups contribute to the overall performance for selected ResNet-50-based methods. Here we train and evaluate reverse probes for each shown combination, *e.g.* IN-1K+OBJECTS, then IN-1K+OBJECTS+SCENE, etc. We compare variants with and without using ground truth IN-1K categories as part of the input. We observe that for earlier methods, such as MoCo-v1, ImageNet categories alone are not sufficient for accurately predicting $f_K(x)$. In fact, it appears that using only semantic categories from MSCOCO and OpenImages (w/o ImageNet) provides about as much information about the clusters. This likely suggests that the discovered clusters reflect fine-grained distinctions, since the most notable difference between IN-1K and the other semantic experts are some fine-grained categories (*e.g.* dog breeds). On the other side of the spectrum, the best performing methods do indeed owe a big part of their performance to IN-1K labels, meaning that they are able to recover clusters that are closer to the original label distribution and that rely less on low-level concepts such as textures. This suggests that methods performing well on linear probing learn features highly correlated with IN-1K labels, but less with other semantic categories.

Finally, in Table 3, we also show the the (decorrelated) effect of each individual attribute group *in isolation* but always in combination with the ground truth IN-1K categories. We observe that

Table 4: Evaluation of selected IN-1k-pretrained methods on the Places-365 dataset. K-means ($K = 500$) and the reverse probe are trained on features extracted on Places-365.

Model	**NMI**	**AMI**	**Top**-1	**mAP**
MoCo-v2	54.41	51.84	41.18	42.15
SwAV	<u>57.30</u>	<u>54.65</u>	44.84	<u>46.48</u>
SeLa-v2	57.02	<u>54.66</u>	<u>45.04</u>	46.42
DeepCluster-v2	**58.09**	**55.62**	**46.23**	**48.06**

in absence of concept combinations, the models we probed use as much (and sometimes more) information about scene and material properties as information about semantic object categories. Miscellaneous concepts (OTHER) do not appear relevant for more recent methods. We again observe that the representations of the most recent, state-of-the-art MoCo-v3 share very little information with concepts other than IN-1K, despite the lack of label information during pre-training.

4.4 QUALITATIVE RESULTS

Since the clusters we examine are learned without supervision and also the expert models are not perfect predictors, it is crucial to verify whether the clusters align with human interpretation, beyond quantitative scores. By inspecting the probe's coefficients θ, one can qualitatively explore what concepts are representative for specific clusters and how exposing the linear model to different sets of experts affects its predictive accuracy. In Fig. 5 we show pairs of clusters from MoCo-v2 for which we observed a significant drop in the confusion (estimated from the confusion matrix) after including an additional concept group on top of the fixed ImageNet label set. For example, after including OBJECTS we notice that several pairs of clusters become more easily disambiguated thanks to the concept `person`, or related concepts (`human hand`, `human face`, etc.). In Fig. 5, since both clusters contain the ImageNet label `French horn`, it is only possible to tell them apart when more information is available. This finding supports our initial hypothesis there could be interpretable properties in the network that may remain undetected since they are not annotated *a-priori* in the benchmark data and further motivates the need for reverse linear probing. Importantly, since ImageNet is only annotated with a single label per image, combined concepts, such as "person playing the french horn" cannot be discovered otherwise. Similarly, when adding SCENE concepts, the two demonstrated clusters are distinguished by the scene type; one contains `volcanos`, while the other contain `mountains` and `creeks` — categories annotated in Places-365. When adding MATERIAL, we can better predict the shipwreck cluster more accurately, while TEXTURE can explain the difference in the more `stratified` appearance of cliffs for the cluster on the right.

4.5 TRANSFER TO OTHER DATASETS

We finally evaluate the representations learned by selected methods by transferring them to a different domain. For this experiment we use Places-365 and note that the self-supervised methods are *not* fine-tuned or adapted on the new data. Instead, we perform the quantization on representations extracted directly on Places-365 with $K = 500$ and report the performance in Table 4. We see that the same ranking generally holds.

5 CONCLUSION

We have introduced reverse linear probing, a measure of representation interpretability that allows to explain representations by combining multiple concepts. Driven by the need to better understand and characterize the representations learned by self-supervised methods, the proposed measure has a rigorous foundation in information theory and provides complementary information to existing benchmarks as well as insights into the semanticity and other interesting properties of such models. The proposed model is also considerably faster to train, as all concepts can be pre-extracted on any given image collection. Code and data will be released upon acceptance.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017. 1, 3

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 2

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 2, 6, 8, 17, 18

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proc. CVPR*, pages 9365–9374, 2019. 15

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Proc. NeurIPS*, 32:9453–9463, 2019. 18

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proc. CVPR*, 2017. 3, 4

Miguel A Bautista, Artsiom Sanakoyeu, Ekaterina Tikhoncheva, and Bjorn Ommer. Cliquecnn: Deep unsupervised exemplar learning. In *NeurIPS*, pages 3846–3854, 2016. 2

Miguel A Bautista, Artsiom Sanakoyeu, and Bjorn Ommer. Deep unsupervised similarity learning using partially ordered sets. In *CVPR*, pages 7130–7139, 2017. 2

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proc. CVPR*, 2015. 5, 15

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018a. 15

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proc. CVPR*, 2018b. 5

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proc. ECCV*, 2018. 2

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Proc. NeurIPS*, 2020. 2, 6, 8

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 2, 6

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 15

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, 2020a. 2, 6

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 2

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b. 2, 6, 8

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 6, 8

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. URL http://www.robots.ox.ac.uk/~vgg/data/dtd/. 5, 15

Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? *arXiv preprint arXiv:2105.05837*, 2021. 3

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2006. 4

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 5

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proc. ICCV*, 2015. 2

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Proc. ICLR*, 2021. 7

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021. 3

Ruth Fong and Andrea Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

Jonathan Frankle, David J Schwab, Ari S Morcos, et al. Are all negatives created equal in contrastive instance discrimination? *arXiv preprint arXiv:2010.06682*, 2020. 2

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *Proc. ICLR*, 2018. 2

Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Perez. Obow: Online bag-of-visual-words generation for self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6840, 2021. 6, 8

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. *arXiv.cs*, abs/1905.01235, 2019a. 6

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6391–6400, 2019b. 3

Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. Vissl. https://github.com/facebookresearch/vissl, 2021. 6

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *Proc. NeurIPS*, 2020. 2, 6

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 6, 15

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, pages 9729–9738, 2020. 2, 6, 8, 17

Olivier J. Hénaff, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv.cs*, abs/1905.09272, 2019. 2

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. ISSN 0036-8075. doi: 10.1126/science.1127647. URL https://science.sciencemag.org/content/313/5786/504. 2

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *ICML*, pages 1558–1567, 2017. 2

Xu Ji, Joao F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 16

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Proc. NeurIPS*, 2020. 2, 6

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020. 3

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 3

Klemen Kotar, Gabriel Ilharco, Ludwig Schmidt, Kiana Ehsani, and Roozbeh Mottaghi. Contrasting contrastive self-supervised representation learning models. *arXiv preprint arXiv:2103.14005*, 2021. 3

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, pages 1–26, 2020. 5

Iro Laina, Ruth C. Fong, and Andrea Vedaldi. Quantifying learnability and describability of visual concepts emerging in representation learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Proc. ECCV*, 2016. 2

Juho Lee, Yoonho Lee, and Yee Whye Teh. Deep amortized clustering. *arXiv preprint arXiv:1909.13433*, 2019. 2

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C.H. Hoi. Prototypical contrastive learning of unsupervised representations. *ICLR*, 2021. 2, 6

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. ECCV*, 2014. 5, 15

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 4

Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016a. 3

Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural preimages. *International Journal of Computer Vision (IJCV)*, 120(3), 2016b. 3

I. Misra and L. van der Maaten. Self-supervised learning of pretext-invariant representations. In *arXiv*, 2020. 6

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proc. CVPR*, pages 6707–6717, 2020. 2

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proc. CVPR*, 2017. 3

Anh Mai Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Proc. NeurIPS*, 2016. 3

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. ECCV*, 2016. 2

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11), 2017. 3

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

Genevieve Patterson and James Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *Proc. CVPR*, 2012. 5, 15

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 16

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proc. NeurIPS*, 2016. 15

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. ImageNet large scale visual recognition challenge. *IJCV*, 115(3), 2015. 5, 6, 15

Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. Concept generalization in visual representation learning. *arXiv preprint arXiv:2012.05649*, 2020. 3

Vikash Sehwag, Mung Chiang, and Prateek Mittal. On separability of self-supervised representations. *ICML workshop on Uncertainty and Robustness in Deep Learning (UDL)*, 2020. 3

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 15

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 3

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv.cs*, abs/1906.05849, 2019. 2, 6

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *Proc. NeurIPS*, 2020. 2, 6

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 6, 7

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Demystifying self-supervised learning: An information-theoretical framework. *arXiv preprint arXiv:2006.05576*, 2020. 2

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *Proc. ICLR*, 2020. 2

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 17, 18

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 6

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12884–12893, 2021. 3

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proc. ICML*, pages 9929–9939. PMLR, 2020. 2

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. CVPR*, 2018. 2

Jia Xue, Hang Zhang, Ko Nishino, and Kristin Dana. Differential viewpoints for ground terrain material recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 15

Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a. 2

Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020b. 6

Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *NeurIPS*, 2020. 3

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 2, 6

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV*, 2014. 3

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 2

Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6688–6697, 2020. 2

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Proc. ECCV*, 2016. 2

Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proc. CVPR*, 2017. 2

Nanxuan Zhao, Zhirong Wu, Rynson W.H. Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? In *Proc. ICLR*, 2021. 3

Evgenii Zheltonozhskii, Chaim Baskin, Alex M Bronstein, and Avi Mendelson. Self-supervised learning for large-scale unsupervised image clustering. *arXiv preprint arXiv:2008.10312*, 2020. 3

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 3

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *PAMI*, 2016. 5

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 15

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. 3

A IMPLEMENTATION DETAILS

We provide all information about the concepts and experts used in our experiments in full detail. Expert models are trained on various tasks and image datasets to extract information about the scene and its objects, material, texture and miscellaneous (other). Then, the trained models are applied to the target dataset (*e.g.* ImageNet (Russakovsky et al., 2015)). While the tasks differ in nature — *e.g.* segmentation (dense), detection, classification — in all cases we convert the predictions to binary vectors. For example, in the case of segmentation and detection, we return only a (binary) label vector denoting whether a class is present in a given image, discarding dense/location information. We concatenate the experts' binary predictions forming the concept vector $y(x)$ for image x, *i.e.*

$$y(x) = [y^{(1)}(x); y^{(2)}(x); \ldots; y^{(m)}(x)],$$

for m expert models with $y^{(i)}$ denoting the i-th expert. In the following, we provide some expert-specific details and summarize them in Table 5.

Detection on Open Images. We use the Tensorflow Object Detection API and publicly available Faster R-CNN model (Ren et al., 2016), trained on Open Images (600 object categories). We apply the model to our target datasets and keep only object categories predicted with a confidence higher than 0.5.

Segmentation on MS COCO. We use a DeepLab-v2 model (Chen et al., 2017) trained for segmentation on MS COCO 2017 (Lin et al., 2014; Caesar et al., 2018a), which includes 91 'thing' and 91 'stuff' categories. 'Thing' categories include common, countable objects with a well-defined shape, such as person, dog, bicycle, etc.; 'stuff' categories typically include background regions, such as sky, snow, wall, grass, etc. and are thus closely related to material properties. While we experimented with confidence and minimum area thresholds for a predicted class to be considered in our scenario, we found that simply returning all predicted categories performed best.

Scene Classification and Scene Attributes. Although ImageNet is an object-centric dataset, a lot of images are related to specific scene types, *e.g.* outdoor scenery (mountains, seaside) or indoor scenery (houses, shops). In fact, some ImageNet categories can even further subdivided by scene type, for example dogs sitting *indoors* or playing in the *park*. It is thus crucial to consider scene classification in our investigations. We use a pre-trained ResNet-50 model provided by the official repository of the Places-365 database (Zhou et al., 2017) and assign each image of our target dataset to a scene category. Further, we use the provided unified model (wide ResNet-18) to predict multiple scene attributes per image. The model is trained on the SUN Attribute Database (Patterson and Hays, 2012), which contains 102 categories describing scene properties, *e.g.* man-made vs. natural or enclosed vs. open area.

Texture and Material Classification. For further detecting material concepts, we train a Deep Encoding Network (DEP) (Xue et al., 2020) on the Materials in Context (MINC) dataset (Bell et al., 2015), which consists of 23 categories (include a background class). We then do the same on the Describable Textures Dataset (DTD) (Cimpoi et al., 2014), which consists of close-up images for 47 texture categories. DEP uses ResNet-50 (He et al., 2016) as backbone and is trained for single-label classification. We apply the trained models to our target datasets, thus returning a single label per image in each case.

Text Detection. Next, we wish to identify the presence of text in the images to investigate whether this is a deciding factor in clustering the image features (*e.g.* camera dates or watermarks). As a text detection expert, we use CRAFT (Baek et al., 2019) (official implementation) without the refinement step. We only return one bit representing whether the image contains text or not, rather than character recognition. Overall, we found that text as a concept did not contribute significantly in our experiments.

Sentiment Analysis. We use a VGG-19 classifier (Simonyan and Zisserman, 2014) trained on the Twitter for Sentiment Analysis Dataset (T4SA) to classify the sentiment for each image into one of three classes (positive/neutral/negative) and use the sentiment prediction as an additional concept. As with text, we found that sentiment does not significantly affect the performance. We found the predictions of the expert model to be somewhat unreliable; this is also likely due to the fact that sentiment is a rather subjective quality.

Table 5: Summarization of expert models and datasets used in our work for concept extraction.

Category	Datasets	#Classes	Task	Model	Source
OBJECTS	Open Images	600	Segmentation	Faster R-CNN	*
	MS COCO (thing)	91	Detection	DeepLab-v2 (ResNet-101)	*
SCENES	Places-365	365	Single-label Classification	ResNet-50	*
	SUN Attribute	102	Multi-label Classification	WideResNet-18	
MATERIAL	MS COCO (stuff)	91	Segmentation	DeepLab-v2 (ResNet-101)	*
	MINC	23	Single-label Classification	DEP (ResNet50)	*
TEXTURE	DTD	47	Single-label Classification	DEP (ResNet-50)	*
	Color	11	Quantization	-	*
OTHER	SynthText, IC13, IC17	1	Text Detection	CRAFT	*
	T4SA	3	Sentiment Classification	VGG-19	*
	OpenAI Data (unreleased)	9	Image/Text Similarity	CLIP	*

Photography. We have observed differences in the photographic style, quality and type of images in ImageNet, *e.g*. ranging from macro photography, usually for flowers or insects, to vector illustrations. In order to probe the representations for style and image quality, we use the recent CLIP model (Radford et al., 2021), trained on 400 million image-text pairs collected from the web and create a set of candidate sentences: "a high quality photo", "a noisy, grainy image", "a blurry image of low quality", "macro photography", "a photo with out of focus background, bokeh effect", "an animated picture, a vector illustration", "a painting", "a portrait". We use the pre-trained CLIP model to compute image and sentence features (normalized to unit norm) and use the dot product to find the similarity between each image and each one of the sentences. We pick the sentence with the highest similarity as the predicted class for each image if the score is at least 0.5, else we assign a background class. Since CLIP operates on free-form text rather than a fixed set of categories, its use is not limited to this type of queries; it can be used as an expert where a labelled dataset might unavailable.

Training details linear model We train K-means for 100 steps using `faiss` (Johnson et al., 2017) and choose the best out of 5 runs. We divide the data into trainval and test sets by splitting all cluster assignments with a 80/20 ratio. We then train the linear model with the cluster assignments as targets and concept vectors as inputs (where the concepts are pre-extracted on a given image collection, *e.g*. ImageNet). We then train the linear model for up to 100 epochs with batch size 512 and optimize using SGD with a momentum of 0.9 and initial learning rate of 3.5 which is further reduced by a factor of 10 at epochs 60 and 80. We also add L2-regularization with weight 3×10^{-6}.

B FURTHER ANALYSIS

In addition to the results and discussion in the main paper, we provide more investigations and insights.

B.1 MUTUAL INFORMATION BETWEEN CONCEPTS AND IMAGENET CATEGORIES

One interesting question that arises is to which extent the elementary concepts we are considering are predictive of ImageNet labels. We answer this question by training a linear model to predict the ground truth label (instead of a pseudo-label) for each image from its concepts. This results in top-1 accuracy of 46.8% (NMI: 64.2, AMI: 54.2) — and significantly less if we exclude 'object' concepts that overlap with ImageNet labels. This suggests that using additional concepts provides information which is *complementary* to the fixed label set of ImageNet, further justifying our approach.

B.2 VARYING K

As discussed in Section 3, the mutual information between a representation $f(x)$ and a fixed set of concepts $y(x)$ is maximized when $f(x) = x$, *i.e*. any processing of x cannot increase the amount of information. It is thus to be expected that with an increasing number of clusters (K), $I(f_K(x), y(x)$ will monotonically increase, as we approximate $f(x)$ (*i.e*. when each sample is its own cluster).

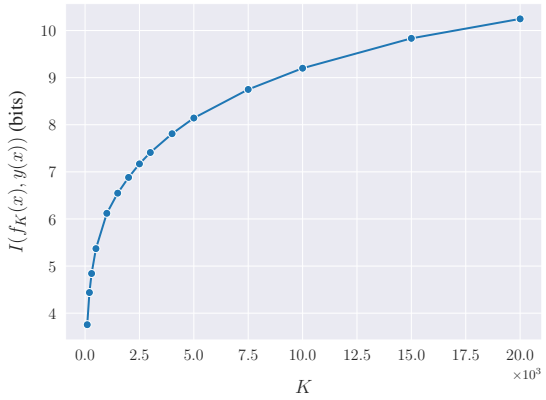


Figure 6: Empirically estimated mutual information between $f_K(x)$ and $y(x)$ for varying K.

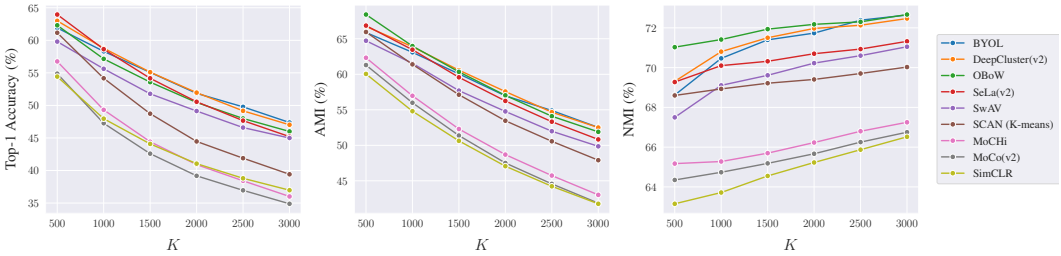


Figure 7: Performance of self-supervised methods for varying number of clusters K. The ranking remains mostly consistent.

We verify this empirically in Fig. 6. For a fair comparison across methods, we have fixed $K = 1000$ for all experiments on ImageNet experiments in the main paper. To better understand the effect of the number of clusters, in Fig. 7 we also show the performance of selected methods for $K \in \{500, 1000, 1500, 2000, 2500, 3000\}$, measuring the top-1 accuracy, NMI and AMI for the predictions of the reverse linear probe trained with all concepts. We observe that for $K \geq 1000$ the relative ranking of methods remains mostly consistent regardless of the number of clusters. One notable exception is BYOL, which performs on par with the top methods (DeepCluster, OBoW) for larger K.

B.3 K-MEANS FOR SELF-LABELLING METHODS

For self-labelling methods SeLa (Asano et al., 2020) and SCAN (Van Gansbeke et al., 2020) we investigate the difference between (a) the pseudo-labels predicted directly by each method and (b) clustering the learned representations with K-means, *i.e.* same as the experiments reported in the main paper. SeLa follows an optimal transport approach that yields clusters of approximately equal size. We found that this weakens the quality of its self-labels but results in stronger clusters after K-means and consequently in improved performance following reverse probing. Since SeLa is originally trained with 3000 classes[1], in Table 6 we compare its pseudo-labels with $K = 3000$-means clustering in Table 6. SCAN builds on top of a representation learning method, *i.e.* MoCo (He et al., 2020) and, with a learnable clustering approach that removes adverse effects of low-level features, it produces more semantic pseudo-labels, improving upon K-means. Thus SCAN (self-label) results in improved performance in Table 6.

[1] https://github.com/yukimasano/self-label

Method		NMI	AMI	Top-1	mAP
SeLa-v1 (K-means)	(Asano et al., 2020)	64.66	37.37	33.05	29.29
SeLa-v1 (self-label)	(Asano et al., 2020)	62.02	31.92	29.22	24.51
SCAN (K-means)	(Van Gansbeke et al., 2020)	69.02	61.56	54.17	60.44
SCAN (self-label)	(Van Gansbeke et al., 2020)	73.76	73.76	62.10	63.63

Table 6: Evaluation of self-labelling methods. "Self-label" denotes clusters which are predicted *directly from the model* and which are used to train the reverse probe. We compare these to the clusters obtained with K-means on the representation vectors (as in the main paper). $K = 3000$ for SeLa and $K = 1000$ for SCAN to match the number of pseudo-labels predicted by the respective models.

Method	Top-1	Top-5	NMI	AMI
MoCo(v2)	40.30 12.67	75.57 32.45	60.14 26.73	50.38 16.53
MoCHi	41.03 12.64	75.62 31.71	59.89 25.66	50.48 15.98
SwAV	45.73 17.71	78.56 43.64	62.75 26.39	53.77 18.72
OBoW	46.92 12.23	78.73 31.72	64.82 27.45	55.75 16.82
DeepCluster(v2)	47.20 19.73	79.90 46.81	63.25 26.84	54.77 19.43
SeLa(v2)	47.62 20.61	80.87 48.83	62.88 25.32	54.91 18.36

Table 7: Reverse probes evaluated on ObjectNet in comparison to ImageNet (in black).

B.4 Testing on ObjectNet

Next we provide an evaluation of selected self-supervised methods using the reversed probes on the challenging test set ObjectNet (Barbu et al., 2019), evaluating the transferability of the quantized representations. Specifically, we use the the reversed probes trained for each method on ImageNet to assign ObjectNet samples to clusters, measuring the success of the classification (accuracy, NMI, AMI) against the corresponding K-means assignments, *i.e.* we measure if we can successfully predict the cluster assignments from predicted concepts, without any further training/adaptation. Here, we use the probes trained with all experts as input, except for the real ImageNet labels to account for ObjectNet categories that do not appear in ImageNet. We notice a significant drop in performance which suggests that a domain gap is indeed present, though the ranking stays again roughly the same (with the exception of OBoW).

C Qualitative Examples

Finally, in Figs. 8, 9, 10 and 11 we present several additional qualitative examples that highlight the usefulness of reverse probing. Similar to Fig. 5 in the main paper, we begin by showing self-supervised clusters which are similar; they often contain images belonging to the same ImageNet categories. As a result, when training a (reverse) linear probe from ImageNet labels to pseudo-labels (cluster assignments), pairs of clusters with overlapping concepts will typically have high confusion (computed from the confusion matrix). We then add a concept group on top of the ImageNet labels and train a second linear probe, *e.g.* in Fig. 9 we include concepts from Places-365 and SUN Attributes. We can then identify pairs of clusters for which the inclusion of these concepts significantly reduced the ambiguity of the mapping, therefore reducing the confusion. Finally, for each pair of clusters we show the difference in linear model's coefficients as a word cloud, with larger differ-

ences denoted by larger font size. In other words, we show the concepts that are most important to distinguish the two clusters (blue for (i) and red for (ii)).

In this way, we were able to discover and understand fine-grained differences between clusters and even problematic cases when considering only ImageNet labels. For example, we find that several ImageNet categories appear in combination with people (*e.g.* musical instruments). The quantized space of self-supervised representations (for all methods) appears to separate the corresponding images based on whether they contain people or not (examples shown in Fig. 8 (a), (c) and (d)). Noteworthy are the clusters in Fig. 8 (d), both of which correspond to the same ImageNet label (`hockey puck`), although they are visually very different. Including additional concepts, such as `person` or `sports equipment`, justifies their distinction into separate clusters and aligns with human intuition — perhaps even more so than assigning them to the same class. In Fig. 9 we show examples where clusters with images from the same ImageNet category differ by scene-related concepts such as flying vs. perched birds, wolves and orcas in different environments and even the inside vs. outside of a building. Material categories (Fig. 10) are helpful to understand environment or context and whether images contain hands (`skin`). Finally, texture (Fig. 11) can be used to tease apart details at a macro level, *e.g.* cut and whole fruits or peacocks with open or closed plumage. Overall, texture becomes less relevant models which are better at semantic discrimination. We also found that other concepts, such as text, sentiment and image quality do not significantly affect the clustering.

To summarize, we have proposed reverse linear probing as a way to understand whether interpretable clusters form in the representation space of self-supervised methods. While we provide a quantifiable measure for this — more interpretable clusters will be predicted with higher accuracy — we also show that we can qualitatively identify how concepts are encoded in the quantized representations through the linear probe's coefficients. We finally show that such concepts are often complementary to ImageNet labels and can in fact *correctly* push clusters which are seemingly semantically close — based on a fixed label set — farther apart (*e.g.* Fig. 9(d)).

Figure 8: Clusters found by unsupervised methods, where each pair (i)-(ii) contains the same (or similar) ImageNet label(s). Confusion is high when probing with ImageNet categories alone, but is significantly reduced after including **Object** concepts from experts trained on COCO-thing and Open Images (OID) categories. The word clouds show the difference in the regressor coefficients for each pair (absolute value is denoted by increasing font size with blue: (i) > (ii) and red: (ii) > (i)).

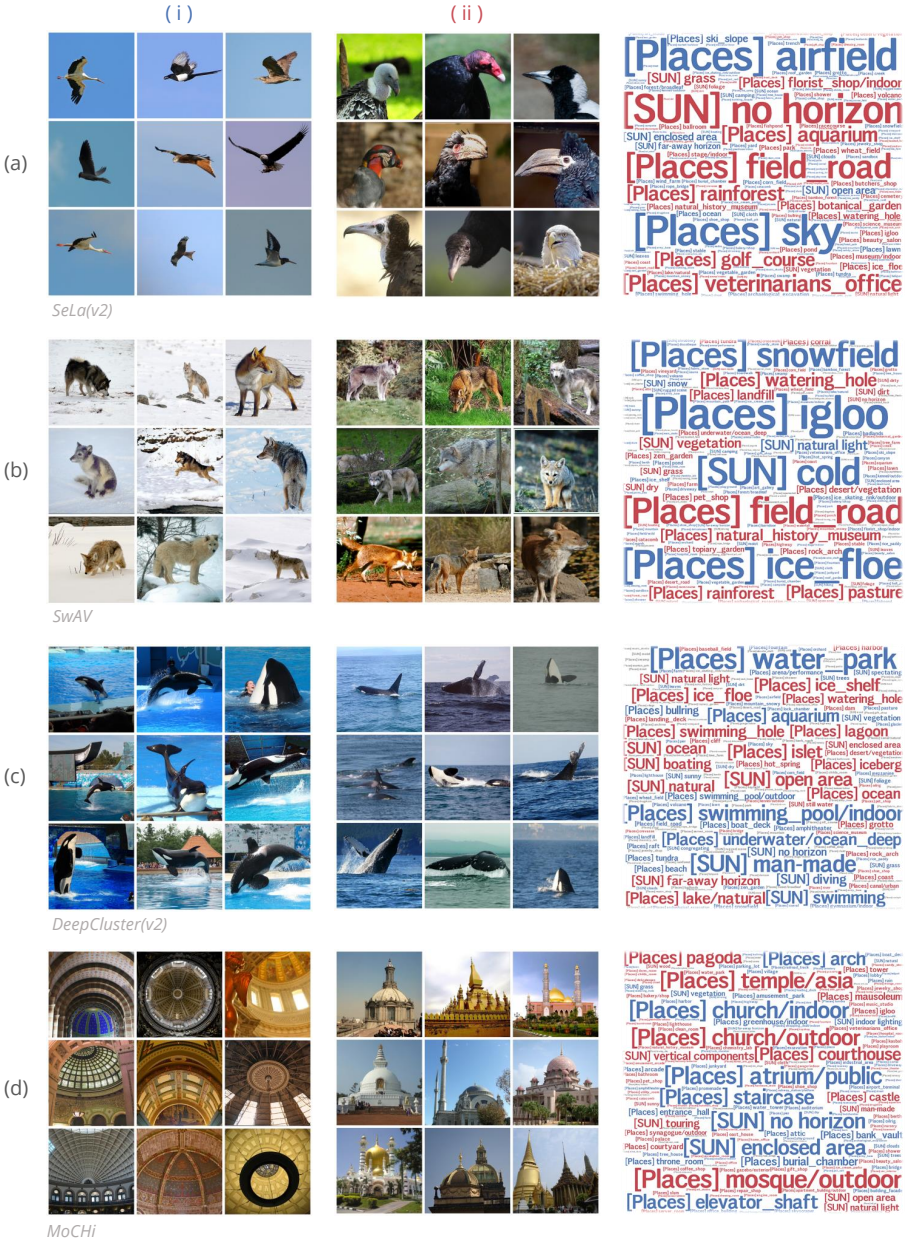


Figure 9: Clusters found by unsupervised methods, where each pair (i)-(ii) contains the same (or similar) ImageNet label(s). Confusion is high when probing with ImageNet categories alone, but is significantly reduced after including **Scene** concepts from experts trained on Places-365 categories and SUN Attributes. The word clouds show the difference in the regressor coefficients for each pair (absolute value is denoted by increasing font size with blue: (i) > (ii) and red: (ii) > (i)).

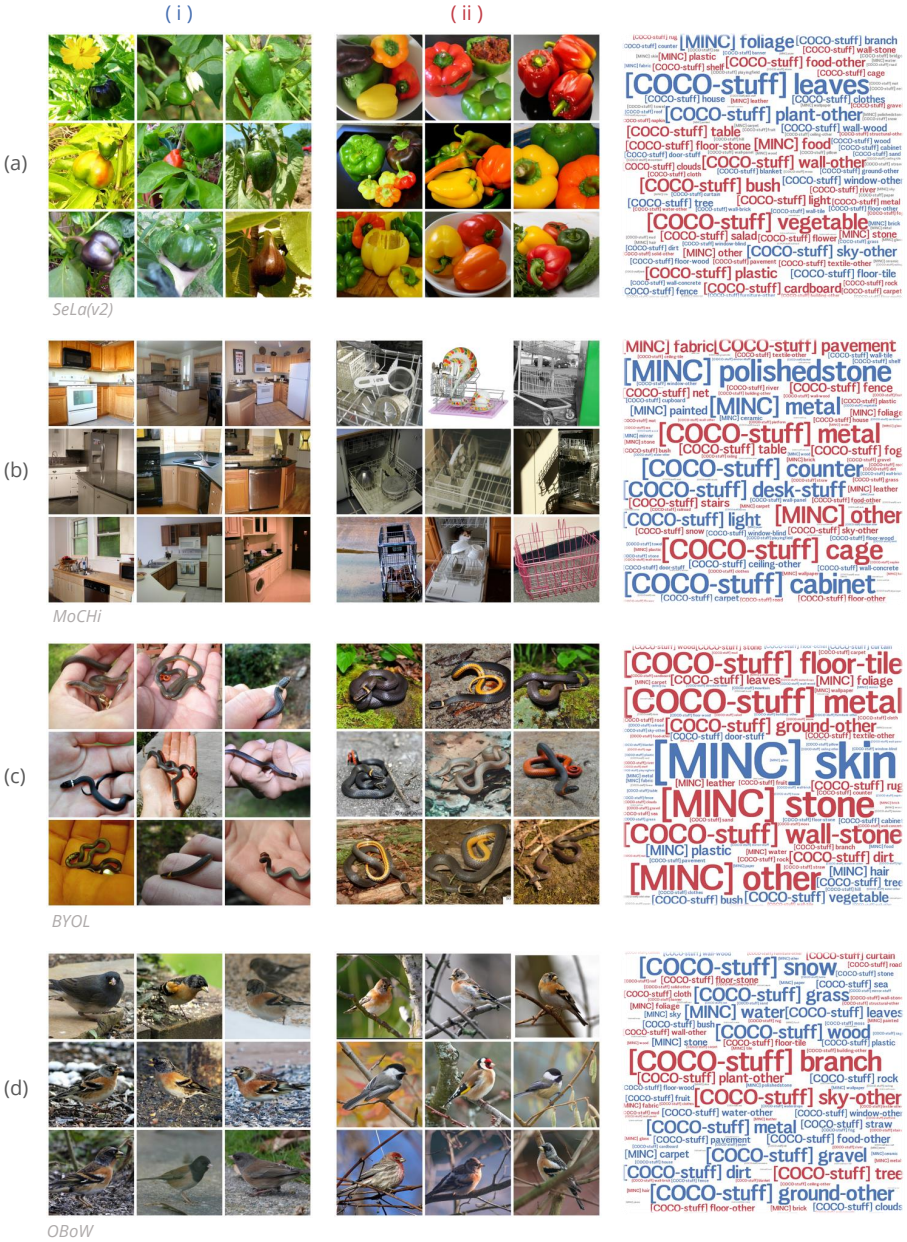


Figure 10: Clusters found by unsupervised methods, where each pair (i)-(ii) contains the same (or similar) ImageNet label(s). Confusion is high when probing with ImageNet categories alone, but is significantly reduced after including **Material** concepts from experts trained on COCO-stuff and Material in Context (MINC). The word clouds show the difference in the regressor coefficients for each pair (absolute value is denoted by increasing font size with blue: (i) > (ii) and red: (ii) > (i)).

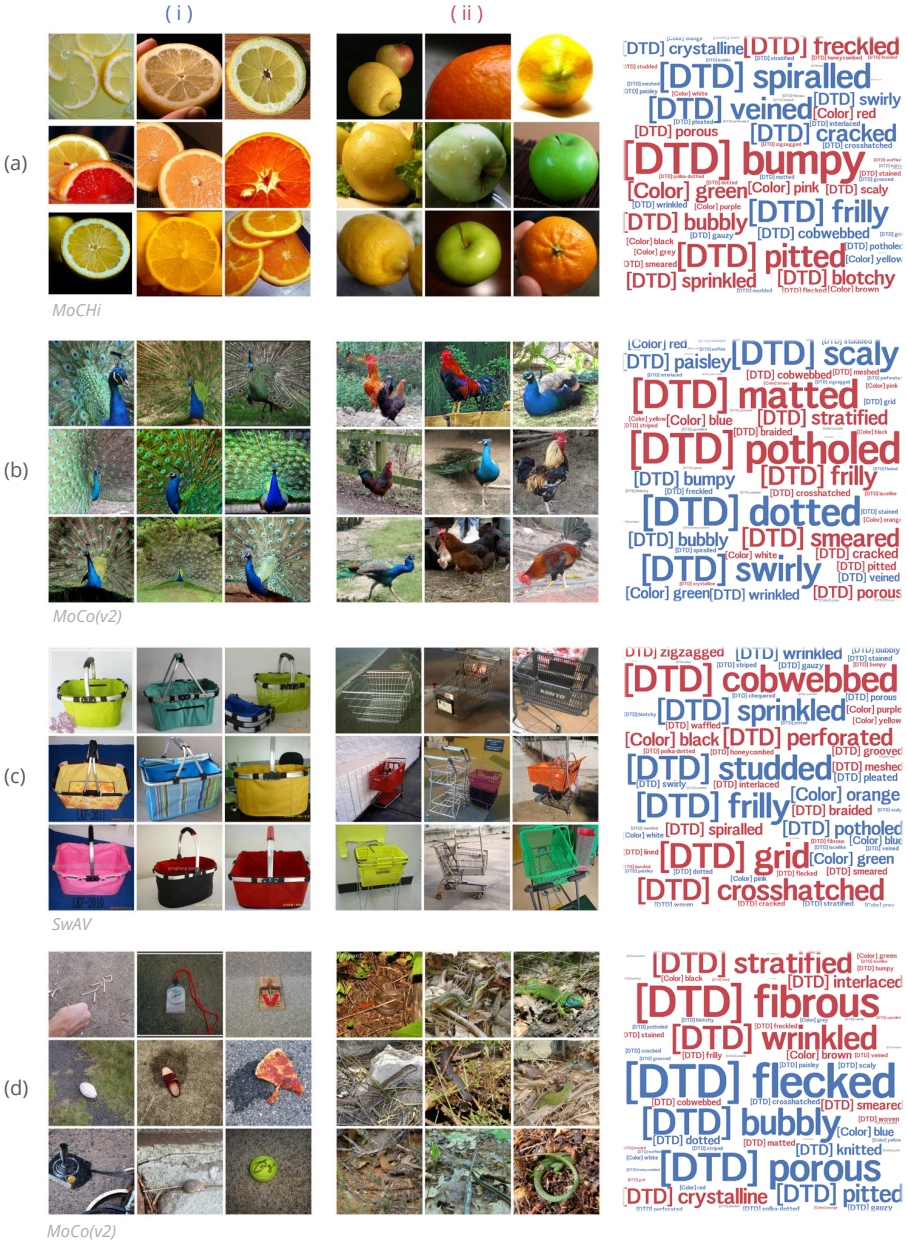


Figure 11: Clusters found by unsupervised methods, where each pair (i)-(ii) contains related ImageNet label(s). Using **Texture** concepts from experts on the Describable Textures Dataset (DTD) and 11 elementary colors helps reduce the confusion comparing to using only ImageNet categories. Clusters where texture helps are usually of lower quality and more recent methods depend less on textures.