

---

# Self-Supervised Relational Reasoning for Representation Learning

---

**Massimiliano Patacchiola**

School of Informatics  
University of Edinburgh  
mpatacch@ed.ac.uk

**Amos Storkey**

School of Informatics  
University of Edinburgh  
a.storkey@ed.ac.uk

## Abstract

In self-supervised learning, a system is tasked with achieving a surrogate objective by defining alternative targets on a set of unlabeled data. The aim is to build useful representations that can be used in downstream tasks, without costly manual annotation. In this work, we propose a novel self-supervised formulation of relational reasoning that allows a learner to bootstrap a signal from information implicit in unlabeled data. Training a relation head to discriminate how entities relate to themselves (*intra-reasoning*) and other entities (*inter-reasoning*), results in rich and descriptive representations in the underlying neural network backbone, which can be used in downstream tasks such as classification and image retrieval. We evaluate the proposed method following a rigorous experimental procedure, using standard datasets, protocols, and backbones. Self-supervised relational reasoning outperforms the best competitor in all conditions by an average 14% in accuracy, and the most recent state-of-the-art model by 3%. We link the effectiveness of the method to the maximization of a Bernoulli log-likelihood, which can be considered as a proxy for maximizing the mutual information, resulting in a more efficient objective with respect to the commonly used contrastive losses.

## 1 Introduction

Learning useful representations from unlabeled data can substantially reduce dependence on costly manual annotation, which is a major limitation in modern deep learning. Toward this end, one solution is to develop learners able to self-generate a supervisory signal exploiting implicit information, an approach known as self-supervised learning (Schmidhuber, 1987, 1990). Humans and animals are naturally equipped with the ability to learn via an intrinsic signal, but how machines can build similar abilities has been material for debate (Lake et al., 2017). A common approach consists of defining a surrogate task (*pretext*) which can be solved by learning generalizable representations, then use those representations in *downstream* tasks, e.g. classification and image retrieval (Jing and Tian, 2020).

A key factor in self-supervised human learning is the acquisition of new knowledge by relating entities, whose positive effects are well established in studies of adult learning (Gentner and Kurtz, 2005; Goldwater et al., 2018). Developmental studies have shown something similar in children, who can build complex taxonomic names when they have the opportunity to compare objects (Gentner and Namy, 1999; Namy and Gentner, 2002). Comparison allows the learner to neglect irrelevant perceptual features and focus on non-obvious properties. Here, we argue that it is possible to exploit a similar mechanism in self-supervised machine learning via relational reasoning.

The relational reasoning paradigm is based on a key design principle: the use of a relation network as a learnable function to quantify the relationships between a set of *objects*. Starting from this principle, we propose a new formulation of relational reasoning which can be used as a pretext task to build useful representations in a neural network backbone, by training the relation head on unlabeled

data. Differently from the canonical relational approach, which focuses on relations between *objects in the same scene* (Santoro et al., 2017), we focus on relations between *views of the same object* (intra-reasoning) and relations between *different objects in different scenes* (inter-reasoning), in doing so we allow the learner to acquire both intra-class and inter-class knowledge without the need of labeled data.

We evaluate our method following a rigorous experimental methodology, since comparing self-supervised learning methods can be problematic (Kolesnikov et al., 2019; Musgrave et al., 2020). Gains may be largely due to the backbone and learning schedule used, rather than the self-supervised component. To neutralize these effects we provide a benchmark environment where all methods are compared using standard datasets (CIFAR-10, CIFAR-100, CIFAR-100-20, STL-10, tiny-ImageNet, SlimageNet), evaluation protocol (Kolesnikov et al., 2019), learning schedule, and backbones (both shallow and deep). Results show that our method largely outperforms the best competitor in all conditions by an average 14% accuracy and the most recent state-of-the-art method by 3%.

*Main contributions:* 1) we propose a novel algorithm based on relational reasoning for the self-supervised learning of visual representations, 2) we show its effectiveness on standard benchmarks with an in-depth experimental analysis, outperforming concurrent state-of-the-art methods (code released with an open-source license<sup>1</sup>), and 3) we highlight how the maximization of a Bernoulli log-likelihood in concert with a relation module, results in more effective and efficient objective functions with respect to the commonly used contrastive losses.

## 1.1 Overview

Following the terminology used in the self-supervised literature (Jing and Tian, 2020) we consider relational reasoning as a *pretext* task for learning useful representations in the underlying neural network backbone. Once the joint system (backbone + relation head) has been trained, the relation head is discarded, and the backbone used in *downstream* tasks (e.g. classification, image retrieval). To achieve this goal we provide a new formulation of relational reasoning. The *canonical formulation* defines it as the process of learning the ways in which entities are connected, using this knowledge to accomplish higher-order goals (Santoro et al., 2017, 2018). The *proposed formulation* defines it as the process of learning the ways entities relate to themselves (intra-reasoning) and to other entities (inter-reasoning), using this knowledge to accomplish downstream goals.

Consider a set of objects  $\mathcal{O} = \{o_1, \dots, o_N\}$ , the canonical approach is *within-scene*, meaning that all the elements in  $\mathcal{O}$  belong to the same scene (e.g. fruits from a basket). The within-scene approach is not very useful in our case. Ideally, we would like our learner to be able to differentiate between objects taken from every possible scene. Therefore first we define *between-scenes* reasoning: the task of relating objects from different scenes (e.g. fruits from different baskets).

Starting from the between-scenes setting, consider the case where the learner is tasked with discriminating if two objects  $\{o_i, o_j\} \sim \mathcal{O}$  belong to the same category  $\{o_i, o_j\} \rightarrow \text{same}$ , or to a different one  $\{o_i, o_j\} \rightarrow \text{different}$ . Often a single attribute is informative enough to solve the task. For instance, in the pair  $\{\text{apple}_i, \text{orange}_j\}$  the color alone is a strong predictor of the class, it follows that the learner does not need to pay attention to other features, this results in poor representations.

To solve the issue we alter the object  $o_i$  via random augmentations  $\mathcal{A}(o_i)$  (e.g. geometric transformation, color distortion) making between-scenes reasoning more complicated. The color of an orange can be randomly changed, or the shape resized, such that it is much more difficult to discriminate it from an apple. In this challenging setting, the learner is forced to take account of the correlation between a wider set of features (e.g. color, size, texture, etc.).

However, it is not possible to create pairs of similar and dissimilar objects when labels are not given. To overcome the problem we bootstrap a supervisory signal directly from the (unlabeled) data, and we do so by introducing *intra-reasoning* and *inter-reasoning*. Intra-reasoning consists of sampling two random augmentations of the same object  $\{\mathcal{A}(o_i), \mathcal{A}(o_i)\} \rightarrow \text{same}$  (positive pair), whereas inter-reasoning consists of coupling two random objects  $\{\mathcal{A}(o_i), \mathcal{A}(o_{\setminus i})\} \rightarrow \text{different}$  (negative pair). This is like coupling different views of the same apple to build the positive pair, and coupling an apple with a random fruit to build the negative pair. In this work we show that it is possible to train a relation module via intra-reasoning and inter-reasoning, with the aim of learning useful representations.

---

<sup>1</sup><https://github.com/mpatacchiola/self-supervised-relational-reasoning>

## 2 Previous work

**Relational reasoning.** In the last decades there have been entire sub-fields interested in relational learning: e.g. reinforcement learning (Džeroski et al., 2001) and statistics (Koller et al., 2007). However, only recently the relational paradigm has gained traction in the deep learning community with applications in question answering (Santoro et al., 2017; Raposo et al., 2017), graphs (Battaglia et al., 2018), sequential streams (Santoro et al., 2018), deep reinforcement learning (Zambaldi et al., 2019), few-shot learning (Sung et al., 2018), and object detection (Hu et al., 2018). Our work differentiate from previous one in several ways: (i) previous work is based on labeled data, while we use relational reasoning on unlabeled data; (ii) previous work has focused on within-scene relations, here we focus on relations between different views of the same object (intra-reasoning) and between different objects in different scenes (inter-reasoning); (iii) in previous work training the relation head was the main goal, here is a pretext task for learning useful representations in the underlying backbone.

**Solving pretext tasks.** There has been a substantial effort in defining self-supervised pretext tasks which can be solved only if generalizable representations have been learned. Examples are: predicting the augmentation applied to a patch (Dosovitskiy et al., 2014), predicting the relative location of patches (Doersch et al., 2015), solving Jigsaw puzzles (Noroozi and Favaro, 2016), learning to count (Noroozi et al., 2017), spotting artifacts (Jenni and Favaro, 2018), predicting image rotations (Gidaris et al., 2018), or image channels (Zhang et al., 2017), generating color version of grayscale images (Zhang et al., 2016; Larsson et al., 2016), and generating missing patches (Pathak et al., 2016).

**Metric learning.** The aim of metric learning (Bromley et al., 1994) is to use a distance metric to bring closer representations of similar inputs (positives), while moving away representations of dissimilar inputs (negatives). Commonly used losses are the contrastive loss (Hadsell et al., 2006), the triplet loss (Weinberger et al., 2006), the Noise-Contrastive Estimation (NCE, Gutmann and Hyvärinen 2010), the margin (Schroff et al., 2015) and magnet (Rippel et al., 2016) losses. At a first glance relational reasoning and metric learning may seem related, however they are fundamentally different: (i) metric learning explicitly aims at organizing representations by similarity, self-supervised relational reasoning aims at learning a relation measure and, as a byproduct, learning useful representations; (ii) metric learning directly applies a distance metric over the representations, relational reasoning collects representations into a set, aggregates them, then estimates relations; (iii) the relational score is not a distance metric (see Section 3.3) but rather a learnable (probabilistic) similarity measure.

**Contrastive learning.** Metric learning methods based on contrastive loss and NCE are often referred to as contrastive learning methods. Contrastive learning via NCE has recently obtained the state of the art in self-supervised learning. However, one limiting factor is that NCE relies on a large quantity of negatives, which are difficult to obtain in mini-batch stochastic optimization. Recent work has used a memory bank to dynamically store negatives during training (Wu et al., 2018), followed by a plethora of other methods (He et al., 2019; Tian et al., 2019; Misra and van der Maaten, 2019; Zhuang et al., 2019). However, a memory bank has several issues, it introduces additional overhead and a considerable memory footprint. SimCLR (Chen et al., 2020) tries to circumvent the problem by mining negatives in-batch, but this requires specialized optimizers to stabilize the training at scale. We compare relational reasoning and contrastive learning in Section 3.1 and Section 5.

**Pseudo-labeling.** Self-supervision can be achieved providing pseudo-labels to the learner, which are then used for standard supervised learning. A way to obtain pseudo-labels is to use the model itself, picking up the class which has the maximum predicted probability (Lee, 2013; Sohn et al., 2020). A neural network ensemble can also be used to provide the labels (Gupta et al., 2020). In DeepCluster (Caron et al., 2018), pseudo-labels are produced by running a k-means clustering algorithm, which can be forced to induce equipartition (Asano et al., 2020). Recent studies have shown that pseudo-labeling is not competitive against other methods (Oliver et al., 2018), since they are often prone to degenerate solutions with points assigned to the same label (or cluster).

**InfoMax.** A recent line of work has investigated the use of mutual information for unsupervised and self-supervised representation learning, following the InfoMax principle (Linsker, 1988). Mutual information is often maximized at different scales (global and local) on single views (Deep InfoMax, Hjelm et al. 2019), multi-views (Bachman et al., 2019; Ji et al., 2019), or sequentially (Oord et al., 2018). Those methods are often strongly dependent on the choice of feature extractor architecture (Tschannen et al., 2020).

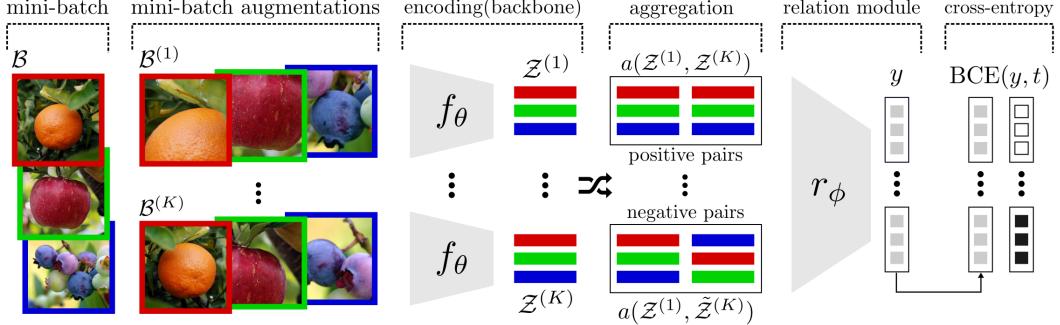


Figure 1: Overview of the proposed method. The mini-batch  $\mathcal{B}$  is augmented  $K$  times (e.g. via random flip and crop-resize) and passed through a neural network backbone  $f_\theta$  to produce the representations  $\mathcal{Z}^{(1)}, \dots, \mathcal{Z}^{(K)}$ . An aggregation function  $a$  joins positives (representations of the same images) and negatives (randomly paired representations) through a commutative operator. The relation module  $r_\phi$  estimates the relational score  $y$ , which must be 1 for positives and 0 for negatives. The model is optimized minimizing the binary cross-entropy (BCE) between prediction and target  $t$ .

### 3 Description of the method

Consider an unlabeled dataset  $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$  and a non-linear function  $f_\theta(\cdot)$  parameterized by a vector of learnable weights  $\theta$ , modeled as a neural network (backbone). A forward pass generates a vector  $f_\theta(\mathbf{x}_n) = \mathbf{z}_n$  (representation), which can be collected in a set  $\mathcal{Z} = \{\mathbf{z}_n\}_{n=1}^N$ . The notation  $\mathcal{A}(\mathbf{x}_n)$  is used to express the probability distribution of instances generated by applying stochastic data augmentation to  $\mathbf{x}_n$ , while  $\mathbf{x}_n^{(i)} \sim \mathcal{A}(\mathbf{x}_n)$  is the  $i$ -th sample from this distribution (a particular augmented version of the input instance), and  $\mathcal{D}^{(i)} = \{\mathbf{x}_n^{(i)}\}_{n=1}^N$  the  $i$ -th set of random augmentations over all instances. Likewise  $\mathbf{z}_n^{(i)} = f_\theta(\mathbf{x}_n^{(i)})$  is grouped in  $\mathcal{Z}^{(i)} = \{\mathbf{z}_n^{(i)}\}_{n=1}^N$ . Let  $K$  indicate the total number of augmentations  $\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(K)}$  and their representations  $\mathcal{Z}^{(1)}, \dots, \mathcal{Z}^{(K)}$ . Now, let us define a relation module  $r_\phi(\cdot)$ , as a non-linear function approximator parameterized by  $\phi$ , which takes as input a pair of aggregated representations and returns a relation score  $y$ . Indicating with  $a(\cdot, \cdot)$  an aggregation function and with  $\mathcal{L}(y, t)$  the loss between the score and a target value  $t$ , the complete learning objective can be specified as

$$\underset{\theta, \phi}{\operatorname{argmin}} \sum_{n=1}^N \sum_{i=1}^K \sum_{j=1}^K \underbrace{\mathcal{L}\left(r_\phi(a(\mathbf{z}_n^{(i)}, \mathbf{z}_n^{(j)})), t=1\right)}_{\text{intra-reasoning}} + \underbrace{\mathcal{L}\left(r_\phi(a(\mathbf{z}_n^{(i)}, \mathbf{z}_{\setminus n}^{(j)})), t=0\right)}_{\text{inter-reasoning}}, \text{ with } \mathbf{z}_n = f_\theta(\mathbf{x}_n), \quad (1)$$

where  $\setminus n$  is an index randomly sampled from  $\{1, \dots, N\} \setminus \{n\}$ . In practice (1) can be framed as a standard binary classification problem (see Section 3.4), and minimized by stochastic gradient descent sampling a mini-batch  $\mathcal{B} \sim \mathcal{D}$  with pairs built by repeatedly applying  $K$  augmentations to  $\mathcal{B}$ . Positives can be obtained pairing two encodings of the same input (intra-reasoning term), and negatives by randomly coupling representations of different inputs (inter-reasoning term), relying on the assumption that in common settings this yields a very low probability of false negatives. An overview of the model is given in Figure 1 and the pseudo-code in Appendix C.

**Mutual information.** Following the recent work of Boudiaf et al. (2020) we can interpret (1) in terms of *mutual information*. Let us define the random variables  $Z|X$  and  $T|Z$ , representing embeddings and targets. Now consider the generative view of mutual information

$$I(Z; T) = H(Z) - H(Z|T). \quad (2)$$

Intra-reasoning is a tightening factor which can be expressed as a bound over the conditional entropy  $H(Z|T)$ . Inter-reasoning is a scattering factor which can be linked to the entropy of the representations  $H(Z)$ . In other words, each representation is pushed towards a positive neighborhood (intra-reasoning) and repelled from a complementary set of negatives (inter-reasoning). Under this interpretation (1) can be considered as a proxy for maximizing Equation (2). We refer the reader to Boudiaf et al. (2020) for a more detailed analysis.

### 3.1 Inputs augmentation

Given a random mini-batch of  $M$  input instances  $\mathcal{B} \sim \mathcal{D}$ , recursively apply data augmentation  $K$  times  $\mathcal{B}^{(1)}, \dots, \mathcal{B}^{(K)}$  then propagate through  $f_\theta$  with a forward pass, to generate the corresponding representations  $\mathcal{Z}^{(1)}, \dots, \mathcal{Z}^{(K)}$ . Representations are coupled across augmentations to generate positive and negative tuples

$$\forall i, j \in \{1, \dots, K\} \quad (\underbrace{\mathcal{Z}^{(i)}, \mathcal{Z}^{(j)}}_{\text{positives}}) \quad \text{and} \quad (\underbrace{\mathcal{Z}^{(i)}, \tilde{\mathcal{Z}}^{(j)}}_{\text{negatives}}), \quad (3)$$

where  $\tilde{\mathcal{Z}}$  indicates random assignment of each representation  $\mathbf{z}_n^{(i)}$  to a different element  $\mathbf{z}_{\setminus n}^{(j)}$ . In practice, we discard identical pairs (identity mapping is learned across augmentations) and take just one of the symmetrical tuples  $(\mathbf{z}^{(i)}, \mathbf{z}^{(j)})$  and  $(\mathbf{z}^{(j)}, \mathbf{z}^{(i)})$  (the aggregation function ensures commutation, see Section 3.2). If a certain amount of data in  $\mathcal{D}$  is labeled (semi-supervised setting), then positive pairs include representations of different augmented inputs belonging to the same category.

**Computational cost.** Having defined  $M$  as the number of inputs in the mini-batch  $\mathcal{B}$ , and  $K$  as the number of augmentations, the total number of pairs  $P$  (positive and negative) is given by

$$P = M(K^2 - K). \quad (4)$$

The number of comparisons  $P$  scales quadratically with the number of augmentations  $K$ , and linearly with the size of the mini-batch  $M$ ; whereas in recent contrastive learning methods (Chen et al., 2020), they scale as  $P = (MK)^2$ , which is quadratic in both augmentations and mini-batch size.

**Augmentation strategy.** Here, we consider the particular case where the input instances are color images. Following previous work (Chen et al., 2020) we focus on two augmentations: random crop-resize and color distortion. Crop-resize enforces comparisons between views: global-to-global, global-to-local, and local-to-local. Since augmentations are sampled from the same color distribution, the color alone may suffice to distinguish positives and negatives. Color distortion enforces color-invariant encodings and neutralizes learning shortcuts. Additional details about the augmentations used in this work are reported in Section 4 and Appendix A.3.

### 3.2 Aggregation function

Relation networks operate over sets. To avoid a combinatorial explosion due to an increasing cardinality, a commutative aggregation function is applied. Given  $f_\theta(\mathbf{x}_i) = \mathbf{z}_i$  and  $f_\theta(\mathbf{x}_j) = \mathbf{z}_j$ , there are different possible choices for the aggregation function

$$a_{\text{sum}}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i + \mathbf{z}_j, \quad a_{\text{max}}(\mathbf{z}_i, \mathbf{z}_j) = \max(\mathbf{z}_i, \mathbf{z}_j), \quad a_{\text{cat}}(\mathbf{z}_i, \mathbf{z}_j) = (\mathbf{z}_i, \mathbf{z}_j), \quad (5)$$

where sum and max are applied elementwise. Concatenation  $a_{\text{cat}}$  is not commutative, but it has been previously used when the cardinality is small (Hu et al., 2018; Sung et al., 2018), like in our case.

### 3.3 Relation module

The relation module is a function  $r_\phi(\cdot)$  parameterized by a vector of learnable weights  $\phi$ , modeled as a multi-layer perceptron (MLP). Given a pair of representations  $\mathbf{z}_i$  and  $\mathbf{z}_j$ , the module takes as input the aggregated pair and produce a scalar  $y$  (relation score)

$$r_\phi(a(\mathbf{z}_i, \mathbf{z}_j)) = y. \quad (6)$$

The relational score respects two properties: (i)  $r(a(\mathbf{z}_i, \mathbf{z}_j)) \in [0, 1]$ ; (ii)  $r(a(\mathbf{z}_i, \mathbf{z}_j)) = r(a(\mathbf{z}_j, \mathbf{z}_i))$ . It is crucial to not misinterpret the relational score for a pairwise distance metric. Given a set of input vectors  $\{\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k\}$  the distance metric  $d(\cdot, \cdot)$  respects four properties: (i)  $d(\mathbf{v}_i, \mathbf{v}_j) \geq 0$ ; (ii)  $d(\mathbf{v}_i, \mathbf{v}_j) = 0 \leftrightarrow \mathbf{v}_i = \mathbf{v}_j$ ; (iii)  $d(\mathbf{v}_i, \mathbf{v}_j) = d(\mathbf{v}_j, \mathbf{v}_i)$ ; (iv)  $d(\mathbf{v}_i, \mathbf{v}_k) \leq d(\mathbf{v}_i, \mathbf{v}_j) + d(\mathbf{v}_j, \mathbf{v}_k)$ . Note that the relational score does not satisfies all the conditions of a distance metric and therefore *the relational score is not a distance metric*, but rather a probabilistic estimate (see Section 3.4).

### 3.4 Definition of the loss

The learning objective (1) can be framed as a binary classification problem over the  $P$  representation pairs. Under this interpretation, the relation score  $y$  represents a probabilistic estimate of representation membership, which can be induced through a sigmoid activation function. It follows that the objective reduces to the maximization of a Bernoulli log-likelihood, or similarly, the minimization of a binary cross-entropy loss

$$\mathcal{L}(\mathbf{y}, \mathbf{t}, \gamma) = \frac{1}{P} \sum_{i=1}^P -w_i \left[ t_i \cdot \log y_i + (1 - t_i) \cdot \log(1 - y_i) \right], \quad (7)$$

with target  $t_i = 1$  for positives and  $t_i = 0$  for negatives. The optional weight  $w_i$  is a scaling factor

$$w_i = \frac{1}{2} \left[ (1 - t_i) \cdot y_i + t_i \cdot (1 - y_i) \right]^\gamma, \quad (8)$$

where  $\gamma \geq 0$  defines how sharp the weight should be. This factor gives more importance to uncertain estimations and it is also known as the focal loss (Lin et al., 2017). Note that, a binary estimator has been previously used in the context of correlation minimization for independent component analysis (Brakel and Bengio, 2017) and in information maximization for representation learning (Hjelm et al., 2019). Hjelm et al. (2019) did not find any major benefit in using a binary loss (the Jensen-Shannon estimator), but similarly to us they observed a low sensitivity to the number of negative samples, outperforming NCE as the number of negatives became smaller (see Section 5 for a discussion).

## 4 Experiments

Evaluating self-supervised methods is problematic because of substantial inconsistency in the way methods have been compared (Kolesnikov et al., 2019; Musgrave et al., 2020). We provide a standardized environment implemented in Pytorch using standard datasets (CIFAR-10, CIFAR-100, CIFAR-100-20, STL-10, tiny-ImageNet, SimageNet), different backbones (shallow and deep), same learning schedule (epochs), and well known evaluation protocols (Kolesnikov et al., 2019). In most conditions our method show superior performance.

**Implementation.** Hyperparameters (relation learner): mini-batch of 64 images ( $K = 16$  for ResNet-32 on tiny-ImageNet,  $K = 25$  for ResNet-34 on STL-10,  $K = 32$  for the rest), Adam optimizer with learning rate  $10^{-3}$ , binary cross-entropy loss with focal factor ( $\gamma = 2$ ). Relation module: MLP with 256 hidden units (batch-norm + leaky-ReLU) and a single output unit (sigmoid). Aggregation: we used concatenation as it showed to be more effective (see Appendix B.8, Table 13). Augmentations: horizontal flip (50% chance), random crop-resize, conversion to grayscale (20% chance), and color jitter (80% chance). Backbones: Conv-4, ResNet-8/32/56 and ResNet-34 (He et al., 2016). Baselines: DeepCluster (Caron et al., 2018), RotationNet (Gidaris et al., 2018), Deep InfoMax (Hjelm et al., 2019), and SimCLR (Chen et al., 2020). Those are recent (hard) baselines, with SimCLR being the current state-of-the-art in self-supervised learning. As upper bound we include the performance of a fully supervised learner (it has access to the labels), and as lower bound a network initialized with random weights, evaluated training only the linear classifier. All results are the average over three random seeds. Additional details in Appendix A.

**Linear evaluation.** We follow the linear evaluation protocol defined by Kolesnikov et al. (2019) training the backbone for 200 epochs using the *unlabeled* training set, and then training for 100 epochs a linear classifier on top of the backbone features (without backpropagation in the backbone weights). The accuracy of this classifier on the test set is considered as the final metric to assess the quality of the representations. Our method largely outperforms other baselines with an accuracy of 46.2% (CIFAR-100) and 30.5% (tiny-Imagenet), which is an improvement of +4.0% and +4.7% over the best competitor (SimCLR), see Table 1. Best results are also obtained with the Conv-4 backbone on all datasets. Only in CIFAR-10/ResNet-32 SimCLR is doing better, with a score of 77% against 75% of our method, see Appendix B.1. In the appendix we report the results on the challenging SimageNet dataset used in few-shot learning (Antoniou et al., 2020): 160 low-resolution images for each one of the 1000 classes in ImageNet. On SimageNet our method has the highest accuracy (15.8%,  $K = 16$ ), being better than RotationNet (7.2%) and SimCLR (14.3%).

**Domain transfer.** We evaluate the performance of all methods in transfer learning by training on the unlabeled CIFAR-10 with linear evaluation on the labeled CIFAR-100 (and viceversa). Our

Table 1: Comparison on various benchmarks. Mean accuracy (percentage) and standard deviation over three runs (ResNet-32). Best results in bold. **Linear Evaluation:** training on unlabeled data and linear evaluation on labeled data. **Domain Transfer:** training on unlabeled CIFAR-10 and linear evaluation on labeled CIFAR-100 ( $10 \rightarrow 100$ ), and viceversa ( $100 \rightarrow 10$ ). **Grain:** training on unlabeled CIFAR-100, linear evaluation on coarse-grained CIFAR-100-20 (20 super-classes). **Finetune:** training on the unlabeled set of STL-10, finetuning on the labeled set (ResNet-34).

Method	Linear Evaluation		Domain Transfer		Grain		Finetune	
	CIFAR-100	tiny-ImgNet	10→100	100→10	CIFAR-100-20	STL-10		
Supervised (upper bound)	65.32±0.22	50.09±0.32	33.98±0.71	71.01±0.44	76.35±0.57	69.82±3.36		
Random Weights (lower bound)	7.65±0.44	3.24±0.43	7.65±0.44	27.47±0.83	16.56±0.48	n/a		
DeepCluster (Caron et al., 2018)	20.44±0.80	11.64±0.21	18.37±0.41	43.39±1.84	29.49±1.36	73.37±0.55		
RotationNet (Gidaris et al., 2018)	29.02±0.18	14.73±0.48	27.02±0.20	52.22±0.70	40.45±0.39	83.29±0.44		
Deep InfoMax (Hjelm et al., 2019)	24.07±0.05	17.51±0.15	23.73±0.04	45.05±0.24	33.92±0.34	76.03±0.37		
SimCLR (Chen et al., 2020)	42.13±0.35	25.79±0.35	36.20±0.16	65.59±0.76	51.88±0.48	89.31±0.14		
<i>Relational Reasoning</i> (ours)	<b>46.17±0.17</b>	<b>30.54±0.42</b>	<b>41.50±0.35</b>	<b>67.81±0.42</b>	<b>52.44±0.47</b>	<b>89.67±0.33</b>		

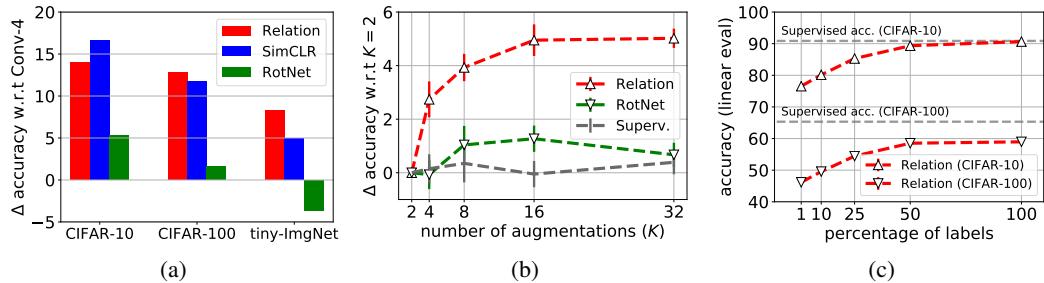


Figure 2: (a) Difference in accuracy using the deeper backbone (Conv4→ResNet-32, linear evaluation). As the complexity of the dataset raises our method performs increasingly better than the others. (b) Correlation between validation accuracy (3 seeds, Conv-4, CIFAR-10) and number of mini-batch augmentations. Only in our method the accuracy is positively correlated with the number of augmentations. (c) Semi-supervised accuracy with an increasing percentage of labels (ResNet-32).

method outperforms once again all the others in every condition. In particular, it is very effective in generalizing from a simple dataset (CIFAR-10) to a complex one (CIFAR-100), obtaining an accuracy of 41.5%, which is a gain of +5.3% over SimCLR and +7.5% over the supervised baseline (with linear transfer). For results see Table 1 and Appendix B.2.

**Grain.** Different methods produce different representations, some may be better on datasets with a small amount of labels (coarse-grained), others may be better on datasets with a large amount of labels (fine-grained). To investigate the granularity of the representations we train on unlabeled CIFAR-100, then perform linear evaluation using the 100 labels (fine grained; e.g. apple, fox, bee, etc) and the 20 super-labels (coarse grained; e.g. fruits, mammals, insects, etc). Also in this case our method is superior in all conditions with an accuracy of 52.4% on CIFAR-100-20, see Table 1 and Appendix B.3. In comparison, the method does better in the fine-grained case, indicating that it is well suited for datasets with a large amount of classes.

**Finetuning.** We used the STL-10 dataset (Coates et al., 2011) which provides a set of unlabeled data coming from a similar but different distribution from the labeled data. Methods have been trained for 300 epochs on the unlabeled set (100K images), finetuned for 20 epochs on the labeled set (5K images), and finally evaluated on the test set (8K images). We used a mini-batch of 64 with  $K = 25$  and a ResNet-34. Implementation details are reported in Appendix A.6. Results in Table 1 show that our method obtains the highest accuracy: 89.67% (best seed 90.04%). Moreover a wider comparison reported in Appendix B.4 shows that the method outperforms strong supervised baselines and the previous self-supervised state-of-the-art (88.80%, Ji et al., 2019).

**Depth of the backbone.** In Appendix B.5 we report an extensive comparison on four backbones of increasing depth: Conv-4, ResNet-8, ResNet-32, and ResNet-56. We tested the three best methods (RotationNet, SimCLR, and Relational Reasoning) on CIFAR-10/100 linear evaluation, grain, and domain transfer for a total of 24 conditions. Results show that our method has the highest accuracy on

21 of those conditions, with SimCLR performing better on CIFAR-10 linear evaluation with ResNet backbones. A distilled version of those results is reported in Figure 2a. The figure shows the gain in accuracy from using a ResNet-32 instead of a Conv-4 backbone for datasets of increasing complexity (10, 100, and 200 classes). As the complexity of the dataset raises our method performs increasingly better than the others. The relative gain against SimCLR gets larger:  $-2.6\%$  (CIFAR-10),  $+1.1\%$  (CIFAR-100),  $+3.3\%$  (tiny-ImageNet). The relative gain against RotationNet is even more evident:  $+8.7\%$ ,  $+11.2\%$ ,  $+11.9\%$ .

**Additional experiments.** Figure 2b and Appendix B.6 show the difference in accuracy between  $K = 2$  and  $K \in \{4, 8, 16, 32\}$  mini-batch augmentations for a fixed mini-batch size. There is a clear positive correlation between the number of augmentations and the performance of our model, while the same does not hold for a self-supervised algorithm (RotationNet) and the supervised baseline. Figure 2c and Appendix B.7 show the accuracy obtained via linear evaluation in the semi-supervised setting, when the number of available labels is gradually increased (0%, 1%, 10%, 25%, 50%, 100%), in both CIFAR-10 and CIFAR-100 (ResNet-32). The accuracy is positively correlated with the proportion of labels available, approaching the supervised upper bound when 100% of labels are available (supervised case).

**Ablations.** In Appendix B.8 we report the results of ablation studies on the aggregation function and relation head. We compare four aggregation functions: sum, mean, maximum, and concatenation. Results show that concatenation and maximum are respectively the most and less effective functions. Concatenation may favor backpropagation improving the quality of the representations, as supported by similar results in previous work (Sung et al., 2018). Ablations of the relation head have followed two directions: (i) removing the head, and (ii) replacing the relation module with an encoder. In the first condition we removed the head and replace it with a simple dot product between representation pairs (BCE-focal loss). In the second condition we followed an approach similar to SimCLR (Chen et al., 2020), replacing the relation head with an encoder and applying the dot product to representations at the higher level (BCE-focal loss). The second condition differs from SimCLR for the loss type (BCE vs Contrastive) and total number of mini-batch augmentations ( $K = 32$  vs  $K = 2$ ). In both conditions we observe a severe degradation of the performance with respect to the complete model (from a minimum of  $-3\%$  to a maximum of  $-23\%$ ), confirming that the relation module is a fundamental component in the pipeline (see discussion in Section 5).

**Qualitative analysis.** In Appendix B.9 is presented a qualitative comparison between the proposed method and RotationNet, on an image retrieval downstream task. Given a random query image (not cherry-picked) the top-10 most similar images in representation space are retrieved. Our method shows better distinction between categories which are hard to separate (e.g. ships vs planes, trucks vs cars). The lower sample variance and the higher similarity with the query, confirm the fine-grained organization of the representations, which account for color, texture, and geometry. An analysis of retrieval errors in Appendix B.10 shows that the proposed method is superior in accuracy across all categories while being more robust against misclassification, with a top-10 retrieval accuracy of 67.8% against 47.7% of RotationNet. In Appendix B.11 we report a qualitative analysis of the representations (ResNet-32, CIFAR-10) using t-SNE (Maaten and Hinton, 2008). Relational reasoning is able to aggregate the data in a more effective way, and to better capture high level relations with lower scattering (e.g. vehicles vs animals super-categories).

## 5 Discussion and conclusions

Self-supervised relational reasoning is effective on a wide range of tasks in both a quantitative and qualitative manner, and with backbones of different size (ResNet-32, ResNet-56 and ResNet-34, with  $0.5 \times 10^6$ ,  $0.9 \times 10^6$  and  $21.3 \times 10^6$  parameters). Representations learned through comparison can be easily transferred across domains, they are fine-grained and compact, which may be due to the direct correlation between accuracy and number of augmentations. An instance is pushed towards a positive neighborhood (intra-reasoning) and repelled from a complementary set of negatives (inter-reasoning). The number of augmentations may have a primary role in this process affecting the quality of the clusters. The possibility to exploit an high number of augmentations, by generating them on the fly, could be decisive in the low-data regime (e.g. unsupervised few-shot/online learning) where self-supervised relational reasoning has the potential to thrive. Those are factors that require further consideration and investigation.

**From self-supervised to supervised.** Recent work has showed that contrastive learning can be used in a supervised setting with competitive results (Khosla et al., 2020). In our experiments we have observed a similar trend, with relational reasoning approaching the supervised performance when all the labels are available. However, we have obtained those results using the same hyperparameters and augmentations used in the self-supervised case, while there may be alternatives that are more effective. Learning by comparison could help in disentangling fine-grained differences in a fully supervised setting with high number of classes, and be decisive to build complex taxonomic representations, as pointed out in cognitive studies (Gentner and Namy, 1999; Namy and Gentner, 2002).

**Comparison with contrastive methods.** We have compared relational reasoning to a state-of-the-art contrastive learning method (SimCLR) using the same backbone, head, augmentation strategy, and learning schedule. Relational reasoning outperforms SimCLR (+3% on average) using a lower number of pairs, being more efficient. Given a mini-batch of size 64, relational reasoning uses  $6.35 \times 10^4$  ( $K = 32$ ) and  $1.5 \times 10^4$  ( $K = 16$ ) pairs, against  $6.55 \times 10^4$  of SimCLR with mini-batch 128. Contrastive losses needs a large number of negatives, which can be gathered by increasing  $M$  the size of the mini-batch, or increasing  $K$  the number of augmentations (both solutions incur a quadratic cost, see Section 3.1). High quality negatives can only be gathered following the first solution, since the second provides lower sample variance. A typical mini-batch in SimCLR encloses 98% negatives and 2% positives, in our method 50% negatives and 50% positives. The larger set of positives could be one of the reasons why relational reasoning is more effective in disentangling fine-grained representations. In addition to the difference in loss type, there is an important structural difference between the two approaches: in SimCLR pairs are allocated in the loss space and then compared via dot product, while in relational reasoning they are aggregated in the space of transferable representations and compared through a relation head. Ablation studies in Section 4 have shown that this structural difference is fundamental for obtaining higher performances, but the way it influences the learning dynamics and the optimization process is not clear and requires further investigation.

**Why does cross-entropy work so well?** We argue that in the context of recent state-of-the-art methods, cross-entropy has been overlooked in favor of contrastive losses. Our experiments show that cross-entropy is a more efficient and effective objective function with respect to the commonly used contrastive losses. Based on the results of the ablation studies, we hypothesize that the difference in performance is mainly due to the use of a relation module in conjunction with the binary cross-entropy loss. When the BCE is split from the relation head and applied directly to the representations there is a drastic drop in performance; applying the BCE to surrogate representations in a second encoding stage (like in SimCLR) is equally ineffective. Therefore, the use of BCE on its own does not provide any advantage but in concert with the relation head it becomes effective. A more thorough analysis is necessary to substantiate these findings, which is left for future work.

## Broader Impact

The motivation behind this work is to build systems able to exploit a large amount of unlabeled data. Applications that could benefit from the proposed method span from standard supervised classifiers to medical diagnostic systems. Therefore, there is a large number of individuals who may benefit or be harmed from this research. This requires putting some effort into selecting the data source, especially when the system is scaled.

In most cases a large body of unlabeled images can be easily gathered from the internet; to avoid biases those images should be representative of different categories. Our method does not guarantee unbiased predictions, therefore it should be used with caution in critical applications. Individuals who may want to use it should consider the particular source of data at hand and evaluate how it could impact the system performance after the final deployment.

## Acknowledgments and Disclosure of Funding

This work was supported by a Huawei DDMPLab Innovation Research Grant.

MP and AS would like to thank anonymous reviewers for useful comments and suggestions; the BayesWatch team for feedback and discussion, in particular Elliot J. Crowley, Luke Darlow, and Joseph Mellor. MP would like to thank the Becchi team for revising the preliminary version of the manuscript, in particular Valerio Biscione, Riccardo Polvara, and Luca Surace.

## References

- Antoniou, A., Patacchiola, M., Ochal, M., and Storkey, A. (2020). Defining benchmarks for continual few-shot learning. *arXiv preprint arXiv:2004.11967*.
- Asano, Y. M., Rupprecht, C., and Vedaldi, A. (2020). Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*.
- Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*.
- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.
- Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P., and Ayed, I. B. (2020). Metric learning: cross-entropy vs. pairwise losses. *arXiv preprint arXiv:2003.08983*.
- Brakel, P. and Bengio, Y. (2017). Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a “siamese” time delay neural network. In *Advances in Neural Information Processing Systems*.
- Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.
- Coates, A., Ng, A., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*.
- DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.
- Doersch, C., Gupta, A., and Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *International Conference on Computer Vision*.
- Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*.
- Džeroski, S., De Raedt, L., and Driessens, K. (2001). Relational reinforcement learning. *Machine learning*, 43(1-2):7–52.
- Gentner, D. and Kurtz, K. (2005). Relational categories. *WK Ahn, RL Goldstone, BC Love, AB Markman, & PW Wolff (Eds.)*, pages 151–175.
- Gentner, D. and Namy, L. L. (1999). Comparison in the development of categories. *Cognitive development*, 14(4):487–513.
- Gidaris, S., Singh, P., and Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*.
- Goldwater, M. B., Don, H. J., Krusche, M. J., and Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General*, 147(1):1.
- Gupta, D., Ramjee, R., Kwatra, N., and Sivathanu, M. (2020). Unsupervised clustering using pseudo-semi-supervised learning. In *International Conference on Learning Representations*.
- Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Conference on Artificial Intelligence and Statistics*.
- Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Computer Vision and Pattern Recognition*.
- Haeusser, P., Plapp, J., Golkov, V., Aljalbout, E., and Cremers, D. (2018). Associative deep clustering: Training a classification network with no labels. In *German Conference on Pattern Recognition*. Springer.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2019). Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*.
- Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2019). Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.
- Hu, H., Gu, J., Zhang, Z., Dai, J., and Wei, Y. (2018). Relation networks for object detection. In *Computer Vision and Pattern Recognition*.

- Jenni, S. and Favaro, P. (2018). Self-supervised feature learning by learning to spot artifacts. In *Computer Vision and Pattern Recognition*.
- Ji, X., Henriques, J. F., and Vedaldi, A. (2019). Invariant information clustering for unsupervised image classification and segmentation. In *International Conference on Computer Vision*.
- Jing, L. and Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.
- Kolesnikov, A., Zhai, X., and Beyer, L. (2019). Revisiting self-supervised visual representation learning. In *Computer Vision and Pattern Recognition*.
- Koller, D., Friedman, N., Džeroski, S., Sutton, C., McCallum, A., Pfeffer, A., Abbeel, P., Wong, M.-F., Heckerman, D., Meek, C., et al. (2007). *Introduction to statistical relational learning*. MIT press.
- Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40.
- Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *European Conference on Computer Vision*. Springer.
- Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *International Conference on Computer Vision*.
- Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3):105–117.
- Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.
- Misra, I. and van der Maaten, L. (2019). Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991*.
- Musgrave, K., Belongie, S., and Lim, S.-N. (2020). A metric learning reality check. *arXiv preprint arXiv:2003.08505*.
- Namy, L. L. and Gentner, D. (2002). Making a silk purse out of two sow’s ears: Young children’s use of comparison in category learning. *Journal of Experimental Psychology: General*, 131(1):5.
- Noroozi, M. and Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*. Springer.
- Noroozi, M., Pirsiavash, H., and Favaro, P. (2017). Representation learning by learning to count. In *International Conference on Computer Vision*.
- Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*.
- Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Oyallon, E., Belilovsky, E., and Zagoruyko, S. (2017). Scaling the scattering transform: Deep hybrid networks. In *International Conference on Computer Vision*.
- Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Computer Vision and Pattern Recognition*.
- Raposo, D., Santoro, A., Barrett, D., Pascanu, R., Lillicrap, T., and Battaglia, P. (2017). Discovering objects and their relations from entangled scene representations. *arXiv preprint arXiv:1702.05068*.
- Rippel, O., Paluri, M., Dollar, P., and Bourdev, L. (2016). Metric learning with adaptive density discrimination. In *International Conference on Learning Representations*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.
- Santoro, A., Faulkner, R., Raposo, D., Rae, J., Chrzanowski, M., Weber, T., Wierstra, D., Vinyals, O., Pascanu, R., and Lillicrap, T. (2018). Relational recurrent neural networks. In *Advances in Neural Information Processing Systems*.
- Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*.

- Schmidhuber, J. (1987). *Evolutionary principles in self-referential learning, or on learning how to learn: The meta-meta-... hook*. Diplomarbeit, Technische Universität München.
- Schmidhuber, J. (1990). Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments.
- Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Computer Vision and Pattern Recognition*.
- Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.
- Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Computer Vision and Pattern Recognition*.
- Tian, Y., Krishnan, D., and Isola, P. (2019). Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.
- Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. (2020). On mutual information maximization for representation learning. In *International Conference on Learning Representations*.
- Weinberger, K. Q., Blitzer, J., and Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*.
- Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Computer Vision and Pattern Recognition*.
- Zambaldi, V., Raposo, D., Santoro, A., Bapst, V., Li, Y., Babuschkin, I., Tuyls, K., Reichert, D., Lillicrap, T., Lockhart, E., et al. (2019). Deep reinforcement learning with relational inductive biases. In *International Conference on Learning Representations*.
- Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. In *European Conference on Computer Vision*. Springer.
- Zhang, R., Isola, P., and Efros, A. A. (2017). Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Computer Vision and Pattern Recognition*.
- Zhuang, C., Zhai, A. L., and Yamins, D. (2019). Local aggregation for unsupervised learning of visual embeddings. In *International Conference on Computer Vision*.

## A Implementation details

### A.1 Datasets

For datasets with low/medium number of categories we used CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), which are composed of  $32 \times 32$  RGB images, with 10 and 100 classes respectively. In addition we used the 20 super-classes of CIFAR-100 (naming this CIFAR-100-20), which consists of broader categories (e.g. fruits, mammals, insects, etc). In the finetuning experiments we used the STL-10 dataset (Coates et al., 2011) which provides 100K RGB images of size  $96 \times 96$  in the unlabeled set, 5K images in the labeled set, and 8K images in the test set.

For datasets with an high number of categories we used the tiny-ImageNet and SimageNet (Antoniou et al., 2020) datasets, both of them derived from ImageNet (Russakovsky et al., 2015). Tiny-ImageNet consists of 200 different categories, with 500 training images ( $64 \times 64$ , 100K in total), 50 validation images (10K in total), and 50 test images (10K in total). SimageNet consists of  $64 \times 64$  RGB images, 1000 categories with 160 training images (160K in total), 20 validation images (20K in total), and 20 test images (20K in total). Both of them are considered more challenging than ImageNet because of the lower resolution of the images and lower number of training samples.

### A.2 Backbones

We use off-the-shelf Pytorch implementations of ResNets as described in the original paper (He et al., 2016). Some of these networks have quite different structure, with ResNet-8/32/56 based on three hyper-blocks (ResNet-32 has  $0.5 \times 10^6$  total parameters) and ResNet-34 based on four hyper-blocks ( $21.3 \times 10^6$  total parameters). The Conv-4 backbone is based on three blocks (8, 16, 32 feature maps), each one performing: convolution (kerne-size=3, stride=1, padding=1), BatchNorm, ReLU, average pooling (kerne-size=2, stride=2). The fourth block (64 feature maps) performed the same operations but with an adaptive average pooling to squeeze the maps to unit shape in the spatial dimension. We used standard fan-in/fan-out weight initialization, and set BatchNorm weights to 1 and bias to 0. For Conv-4 and ResNet-8/32/56 the size of the representations is 64, whereas for ResNet-34 is 512.

### A.3 Augmentations

During the self-supervised training phase of our method we used a set of augmentations which is similar to the one adopted by Chen et al. (2020). We apply horizontal flip (50% chance), random crop-resize, conversion to grayscale (20% chance), and color jitter (80% chance). Random crop-resize consists of cropping the given image (from 0.08 to 1.0 of the original size), changing the aspect ratio (from 3/4 to 4/3 of the original aspect ratio), and finally resizing to input shape using a bilinear interpolation. Color jitter consists of sampling from a uniform distribution  $[0, max]$  a jittering value for: brightness ( $max = 0.8$ ), contrast ( $max = 0.8$ ), saturation ( $max = 0.8$ ), and hue ( $max = 0.2$ ).

### A.4 Computing infrastructure

All the experiments have been performed on a workstation with 20 cores, 187 GB of RAM, and with 8 NVIDIA GeForce RTX 2080 Ti GPUs (11 GB of internal RAM). All the methods could fit on a single one of those GPUs.

### A.5 Other methods

**Supervised.** This baseline consists of standard supervised training. We used standard data augmentation (horizontal flip and random crop) and learning schedule (SGD optimizer with initial learning rate of 0.1 divided by 10 at 50% and 75% of total epochs). It represents an upper bound. When evaluated for the number of augmentations (Appendix B.6) the same strategy adopted in our method (Appendix A.3) has been used to augment the input mini-batch (size 128)  $K$  times with coherent labels.

**Random weights.** This baseline consists of initializing the weights of the backbone via standard fan-in/fan-out, then perform linear evaluation optimizing the last linear layer (without backpropagation on the backbone). It represents a lower bound since the backbone is not trained.

**DeepCluster (Caron et al., 2018).** We adapted the open-source implementation provided by the authors<sup>2</sup>. Clustering has been performed at the beginning of each epoch by using the k-means algorithm available in Scikit-learn. We performed whitening over the features before the clustering step, as suggested by the authors. We used a number of cluster one order of magnitude larger than the number of classes in the dataset, as recommended by the authors to improve the performance. We also used an MLP head (256 hidden units with leaky-ReLU

<sup>2</sup><https://github.com/facebookresearch/deepcluster>

and BatchNorm) instead of a linear layer, since in our tests this showed to slightly boost the performance. The MLP weights have been reset at the beginning of each epoch as in the original code. We optimized the model minimizing the cross-entropy loss between the pseudo-labels provided by the clustering and the network outputs. We used Adam optimizer with learning rate  $10^{-3}$ .

**RotationNet (Gidaris et al., 2018).** Given the simplicity of the method, this has been reproduced locally following the instructions of the authors. Labels are provided by 4 rotations ( $0^\circ, 90^\circ, 180^\circ, 270^\circ$ ), those are the one providing the highest accuracy according to the authors. The input mini-batch of size 128, has been augmented adding 4 rotations for each image (resulting in a mini-batch of size  $128 \times 4$ ). This is in line with the best performing strategy reported by the authors. In all experiments the cross-entropy loss between the network output and the labels provided by the rotation has been minimized (Adam optimizer, learning rate  $10^{-3}$ ). When evaluated for the number of augmentations (Appendix B.6) the same strategy used in our method has been applied (Appendix A.3), augmenting the input mini-batch (size 128)  $K$  times with coherent self-supervised rotation labels. In order to keep the size of the mini-batch manageable the additional 4 rotations for image have not been included, since this would increase the size to  $4 \times K$  and not fit on the available hardware.

**Deep InfoMax (Hjelm et al., 2019)** The code has been adapted from open-source implementations available online (see code for details) and from the code provided by the authors<sup>3</sup>. The local version of the algorithm has been used ( $\alpha = 0, \beta = 1.0, \gamma = 0.1$ ), as reported by the authors this is the one with the best performance. The capacity of the discriminator networks has been partially reduced to fit the available hardware and to speedup the training, this did not affected significantly the results. We used Adam optimizer with learning rate  $10^{-4}$  as in the original paper.

**SimCLR (Chen et al., 2020)** The code has been adapted from the implementation provided by the authors<sup>4</sup> and other open-source implementations (see code for details). To have a fair comparison with our method we used the same MLP head, the same data augmentation strategy, and optimizer (Adam with learning rate  $10^{-3}$ ). We used a temperature of 0.5 in all the experiments, this was reported as the consistent optimal value regardless of the batch sizes in the original paper. We could not replicate the original setup reported by the authors on very large mini-batches, since it is computationally expensive, requiring 32 to 128 cores on Tensor Processing Units (TPUs). We adapted the setup to our available hardware (described in Appendix A.4), and we guaranteed a fair comparison by using a comparable number of pairs. In particular, we used a mini-batch of 128 images, which results in  $6.5 \times 10^4$  pairs, this is similar (or superior) to the number of pairs compared by our method which is  $6.3 \times 10^4$  for  $K = 32$ ,  $3.8 \times 10^4$  for  $K = 25$ , and  $1.5 \times 10^4$  for  $K = 16$ .

## A.6 Finetuning experiments

All methods are trained for 300 epochs on the unlabeled portion of the STL-10 dataset, using the same hyperparameters and augmentations described before and a ResNet34 backbone. In the finetuning stage the pretrained backbone is coupled with a linear classifier and both of them are trained using Adam optimizer for 100 epochs with mini-batch of size 32. A lower learning rate for the backbone ( $10^{-4}$ ) respect to the linear classifier ( $10^{-3}$ ) has been used. Both learning rates are divided by 10 at 50% and 75% of total epochs. The same augmentations of Ji et al. (2019) have been used for the finetuning stage. Those consists of affine transformations (50% chance) sampled from a uniform distribution [ $min, max$ ]: random rotation ( $min = -18^\circ, max = 18^\circ$ ), scale ( $min = 0.9, max = 1.1$ ), translation ( $min = 0, max = 0.1$ ), shear ( $min = -10, max = 10$ ), and bilinear interpolation, cutout (50% chance) with  $32 \times 32$  patches. Same schedule and augmentations have also been used to train (from scratch) the supervised baseline on the labeled set of data (100 epochs).

## A.7 Semi-supervised experiments

We adapted our method to the semi-supervised case by coupling instances sampled from the same category. Those instances represented a portion of the total number of pairs in the mini-batch depending on the percentage of available labels. Results for each conditions are the average of three seeds. We used the same hyperparameters described in the linear evaluation phase. We did not prevent possible collisions between classes during the allocation of negative pairs. Collisions are unlikely in datasets with medium/high number of classes, but a slight performance improvement could be obtained if negatives are paired without collisions.

## A.8 Qualitative analysis experiments

For the qualitative analysis we compared the representations generated by the supervised baseline, Rotation Net, and our method on CIFAR-10 with a ResNet-32 backbone at the end of the training (200 epochs). The query images were randomly sampled and the representations compared using Euclidean distance. For the t-SNE analysis we used the Scikit implementation of the algorithms and set the hyperparameters as follows: 1000 iterations, Euclidean metric, random init, perplexity 30, learning rate 200.

---

<sup>3</sup><https://github.com/rdevon/DIM>

<sup>4</sup><https://github.com/google-research/simclr>

## B Additional results

### B.1 Linear evaluation

Table 2: Linear evaluation. Self-supervised training on unlabeled data and linear evaluation on labeled data. Comparison between three datasets (CIFAR-10, CIFAR-100, tiny-ImageNet) for a shallow (Conv-4) and a deep (ResNet-32) backbone. Mean accuracy (percentage) and standard deviation over three runs. Best results highlighted in bold.

Method	Conv-4			ResNet-32		
	CIFAR-10	CIFAR-100	tiny-ImageNet	CIFAR-10	CIFAR-100	tiny-ImageNet
Supervised (upper bound)	80.46±0.39	49.29±0.85	36.47±0.36	90.87±0.41	65.32±0.22	50.09±0.32
Random Weights (lower bound)	32.92±1.88	10.79±0.59	6.19±0.13	27.47±0.83	7.65±0.44	3.24±0.43
DeepCluster (Caron et al., 2018)	42.88±0.21	21.03±1.56	12.60±1.23	43.31±0.62	20.44±0.80	11.64±0.21
RotationNet (Gidaris et al., 2018)	56.73±1.71	27.45±0.80	18.40±0.95	62.00±0.79	29.02±0.18	14.73±0.48
Deep InfoMax (Hjelm et al., 2019)	44.60±0.27	22.74±0.21	14.19±0.13	47.13±0.45	24.07±0.05	17.51±0.15
SimCLR (Chen et al., 2020)	60.43±0.26	30.45±0.41	20.90±0.15	<b>77.02±0.64</b>	42.13±0.35	25.79±0.40
<i>Relational Reasoning</i> (ours)	<b>61.03±0.23</b>	<b>33.38±1.02</b>	<b>22.31±0.19</b>	74.99±0.07	<b>46.17±0.16</b>	<b>30.54±0.42</b>

Table 3: Linear evaluation on SlimageNet (Antoniou et al., 2020). This dataset is more challenging than ImageNet, since it only has 160 low-resolution ( $64 \times 64$ ) color images for each one of the 1000 classes of ImageNet. Below is reported the linear evaluation accuracy on labeled data with a ResNet-32 backbone, after training on unlabeled data. Mean accuracy (percentage) and standard deviation over three runs. Best result highlighted in bold.

Method	SlimageNet
Supervised (upper bound)	33.94±0.21
Random Weights (lower bound)	0.79±0.09
RotationNet (Gidaris et al., 2018)	7.25±0.28
SimCLR (Chen et al., 2020)	14.32±0.24
<i>Relational Reasoning</i> (ours)	<b>15.81±0.72</b>

### B.2 Domain transfer

Table 4: Domain transfer. Training with self-supervision on unlabeled CIFAR-10 linear evaluation on CIFAR-100 ( $10 \rightarrow 100$ ), and viceversa ( $100 \rightarrow 10$ ).  $\Delta$  indicates the difference between the accuracy in the standard setting (unsupervised train and linear evaluation on the same dataset) and the accuracy in the transfer setting (unsupervised train on first dataset and linear evaluation on the second dataset). Mean accuracy (percentage) and standard deviation over three runs. Best results highlighted in bold.

Method	Conv-4				ResNet-32			
	10 → 100	$\Delta$	100 → 10	$\Delta$	10 → 100	$\Delta$	100 → 10	$\Delta$
Supervised (upper bound)	32.06±0.63	-17.23	64.00±1.07	-16.46	33.98±0.70	-31.34	71.01±0.44	-19.86
Random Weights (lower bound)	10.79±0.59	n/a	32.92±1.89	n/a	7.65±0.44	n/a	27.47±0.83	n/a
DeepCluster (Caron et al., 2018)	19.68±1.23	-1.35	43.59±1.31	+0.71	18.37±0.41	-2.07	43.39±1.84	+0.08
RotationNet (Gidaris et al., 2018)	26.06±0.09	-1.39	51.86±0.36	-4.87	27.02±0.20	-2.00	52.22±0.70	-9.78
Deep InfoMax (Hjelm et al., 2019)	22.35±0.12	-0.39	43.30±0.15	-1.30	23.73±0.04	-0.34	45.05±0.24	-2.08
SimCLR (Chen et al., 2020)	29.20±0.08	-1.25	54.73±0.60	-5.70	36.21±0.16	-5.92	65.59±0.76	-11.43
<i>Relational Reasoning</i> (ours)	<b>31.84±0.23</b>	-1.54	<b>57.30±0.26</b>	-3.73	<b>41.50±0.35</b>	-4.67	<b>67.81±0.42</b>	-7.18

### B.3 Grain

Table 5: Grain. Training with self-supervision on unlabeled CIFAR-100, and linear evaluation on labeled CIFAR-100 Fine-Grained (100 classes) and CIFAR-100-20 Coarse-Grained (20 super-classes). Mean accuracy (percentage) and standard deviation over three runs. Best results highlighted in bold.

Method	Conv-4		ResNet-32	
	Fine-Grain	Coarse-Grain	Fine-Grain	Coarse-Grain
Supervised (upper bound)	49.29±0.85	59.91±0.62	65.32±0.22	76.35±0.57
Random Weights (lower bound)	10.79±0.59	19.94±0.31	7.65±0.44	16.56±0.48
DeepCluster (Caron et al., 2018)	21.03±1.56	30.07±2.06	20.44±0.80	29.49±1.36
RotationNet (Gidaris et al., 2018)	27.45±0.80	35.49±0.17	29.02±0.19	40.45±0.39
Deep InfoMax (Hjelm et al., 2019)	22.74±0.21	32.36±0.43	24.07±0.05	33.92±0.34
SimCLR (Chen et al., 2020)	30.45±0.41	37.72±0.14	42.13±0.35	51.88±0.48
<i>Relational Reasoning</i> (ours)	<b>33.38±1.02</b>	<b>40.86±1.03</b>	<b>46.17±0.17</b>	<b>52.44±0.47</b>

### B.4 Finetuning

Table 6: Finetuning. Comparison with other results reported in the literature on unsupervised training and finetuning on the STL-10 dataset. Best result in bold. *Local* refers to our local reproduction of the method, with results reported as *best (mean ± std)* on three runs with different seeds. Note that backbone and learning schedule may differ. The ResNet-34 backbone is much larger than ResNet-32 ( $21.3 \times 10^6$  vs  $0.47 \times 10^6$ ), showing that the proposed method can be effectively scaled.

Method	Reference	Backbone	Accuracy
Supervised (crop + cutout)	DeVries and Taylor (2017)	WideResnet-16-8	87.30
Supervised (scattering)	Oyallon et al. (2017)	Hybrid-WideResnet	87.60
Exemplars (Dosovitskiy et al., 2014)	Dosovitskiy et al. (2014)	Conv-3	72.80
Artifacts (Jenni and Favaro, 2018)	Jenni and Favaro (2018)	Custom	80.10
ADC (Haeusser et al., 2018)	Ji et al. (2019)	ResNet-34	56.70
DeepCluster (Caron et al., 2018)	Ji et al. (2019)	ResNet-34	73.40
Deep InfoMax (Hjelm et al., 2019)	Ji et al. (2019)	AlexNet	77.00
Invariant Info Clustering (Ji et al., 2019)	Ji et al. (2019)	ResNet-34	88.80
Supervised (affine + cutout)	Local	ResNet-34	72.04 (69.82 ± 3.36)
DeepCluster (Caron et al., 2018)	Local	ResNet-34	74.00 (73.37 ± 0.55)
RotationNet (Gidaris et al., 2018)	Local	ResNet-34	83.77 (83.29 ± 0.44)
Deep InfoMax (Hjelm et al., 2019)	Local	ResNet-34	76.45 (76.03 ± 0.37)
SimCLR (Chen et al., 2020)	Local	ResNet-34	89.44 (89.31 ± 0.14)
<i>Relational Reasoning</i> (ours)	Local	ResNet-34	<b>90.04 (89.67 ± 0.33)</b>

### B.5 Performance with different backbones

Table 7: Comparison on different backbones: linear evaluation. Comparison between four backbones of different depth for baselines and the three best performing methods. Training with self-supervision on unlabeled CIFAR-10 and CIFAR-100, and linear evaluation on labeled version of the same datasets. Mean accuracy (percentage) and standard deviation over three runs. Best results highlighted in bold.

Method	Conv-4		ResNet-8		ResNet-32		ResNet-56	
	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100
Supervised (upper bound)	80.46±0.39	49.29±0.85	87.08±0.17	59.41±1.15	90.87±0.41	65.32±0.22	91.40±0.30	67.54±0.32
Random Weights (lower bound)	32.92±1.89	10.79±0.59	35.94±1.39	13.08±0.91	27.47±0.83	7.65±0.44	13.53±3.66	1.88±0.14
RotationNet (Gidaris et al., 2018)	56.73±1.71	27.45±0.80	62.73±0.94	32.09±0.87	62.00±0.79	29.02±0.18	61.66±1.11	28.24±0.23
SimCLR (Chen et al., 2020)	60.43±0.26	30.45±0.41	<b>69.85±0.58</b>	36.23±0.15	<b>77.02±0.64</b>	42.13±0.35	<b>78.75±0.24</b>	44.33±0.48
<i>Relational Reasoning</i> (ours)	<b>61.03±0.23</b>	<b>33.38±1.02</b>	67.97±0.58	<b>38.18±0.63</b>	74.99±0.07	<b>46.17±0.17</b>	77.51±0.00	<b>47.90±0.27</b>

Table 8: Comparison on different backbones: grain. Comparison between four backbones of different depth for baselines and the three best performing methods. Training with self-supervision on unlabeled CIFAR-100 and linear evaluation on labeled version of the same datasets with 100 labels (fine) or 20 super-labels (coarse). Mean accuracy (percentage) and standard deviation over three runs. Best results highlighted in bold.

Method	Conv-4		ResNet-8		ResNet-32		ResNet-56	
	Fine	Coarse	Fine	Coarse	Fine	Coarse	Fine	Coarse
Supervised (upper bound)	49.29±0.85	59.91±0.62	59.41±1.15	70.12±0.33	65.32±0.22	76.35±0.57	67.54±0.32	77.60±0.43
Random Weights (lower bound)	10.79±0.59	19.94±0.31	13.08±0.91	23.12±0.90	7.65±0.44	16.56±0.48	1.88±0.14	6.88±0.35
RotationNet (Gidaris et al., 2018)	27.45±0.80	35.49±0.17	32.09±0.87	41.21±0.94	29.02±0.18	40.45±0.39	28.24±0.23	39.16±0.35
SimCLR (Chen et al., 2020)	30.45±0.41	37.72±0.14	36.23±0.15	43.78±0.92	42.13±0.35	51.87±0.48	44.33±0.48	54.09±0.15
<i>Relational Reasoning</i> (ours)	<b>33.38±1.02</b>	<b>40.86±1.03</b>	<b>38.18±0.63</b>	<b>45.36±0.55</b>	<b>46.17±0.17</b>	<b>52.44±0.47</b>	<b>47.90±0.27</b>	<b>54.90±0.07</b>

Table 9: Comparison on different backbones: domain transfer. Comparison between four backbones of different depth for baselines and the three best performing methods. Training with self-supervision on unlabeled CIFAR-10 linear evaluation on CIFAR-100 ( $10 \rightarrow 100$ ), and viceversa ( $100 \rightarrow 10$ ). Mean accuracy (percentage) and standard deviation over three runs. Best results highlighted in bold.

Method	Conv-4		ResNet-8		ResNet-32		ResNet-56	
	10→100	100→10	10→100	100→10	10→100	100→10	10→100	100→10
Supervised (upper bound)	32.06±0.63	64.00±1.07	36.83±0.36	71.20±0.18	33.98±0.70	71.01±0.44	33.92±0.50	71.97±0.17
Random Weights (lower bound)	10.79±0.59	32.92±1.89	13.08±0.91	35.94±1.39	7.65±0.44	27.47±0.83	1.88±0.14	13.53±3.66
RotationNet (Gidaris et al., 2018)	26.06±0.09	51.86±0.36	31.60±0.54	56.85±0.13	27.02±0.20	52.22±0.70	27.25±0.62	51.82±0.58
SimCLR (Chen et al., 2020)	29.20±0.08	54.73±0.60	34.46±0.78	61.34±0.24	36.21±0.16	65.59±0.76	36.79±0.45	66.19±0.80
<i>Relational Reasoning</i> (ours)	<b>31.84±0.23</b>	<b>57.30±0.26</b>	<b>36.07±0.35</b>	<b>63.24±0.52</b>	<b>41.50±0.35</b>	<b>67.81±0.42</b>	<b>42.19±0.28</b>	<b>68.66±0.21</b>

## B.6 Number of augmentations

Table 10: Accuracy with respect to the number of augmentations  $K$ . Methods have been trained on CIFAR-10 with a Conv-4 backbone for 100 epochs. The input mini-batch has been augmented  $K$  times then given as input. Results are the average accuracy (linear evaluation) of three runs on the validation set. Only the relational reasoning accuracy is positively correlated with  $K$ .

Method	$K = 2$	$K = 4$	$K = 8$	$K = 16$	$K = 32$
Supervised	79.61±0.47	79.76±0.54	79.96±0.71	79.56±0.49	80.00±0.45
RotationNet (Gidaris et al., 2018)	51.58±0.49	51.51±1.02	52.62±0.68	52.85±1.24	52.25±1.06
<i>Relational Reasoning</i> (ours)	55.31±0.58	58.05±0.67	59.24±0.51	60.26±0.59	60.33±0.36

## B.7 Semi-supervised and supervised

Table 11: Test accuracy on *CIFAR-10* with respect to the percentage of labeled data available. Methods have been trained with a ResNet-32 backbone (200 epochs), followed by linear evaluation on the entire labeled dataset (100 epochs). The quality of the representations improves with the number of labeled data available.

Method	0%	1%	10%	25%	50%	100%
Supervised	n/a	n/a	n/a	n/a	n/a	90.87±0.41
<i>Relational Reasoning</i> (ours)	74.99±0.07	76.55±0.27	80.14±0.35	85.30±0.28	89.35±0.11	90.66±0.23

Table 12: Test accuracy on *CIFAR-100* with respect to the percentage of labeled data available. Methods have been trained with a ResNet-32 backbone (200 epochs), followed by linear evaluation on the entire labeled dataset (100 epochs). The quality of the representations improves with the number of labeled data available.

Method	0%	1%	10%	25%	50%	100%
Supervised	n/a	n/a	n/a	n/a	n/a	65.32±0.22
<i>Relational Reasoning</i> (ours)	46.17±0.17	46.10±0.29	49.55±0.36	54.44±0.58	58.52±0.70	58.96±0.28

## B.8 Ablations

Table 13: Ablation of the aggregation function. Training with relational self-supervision on unlabeled CIFAR-10 and CIFAR-100, and linear evaluation on labeled datasets (Conv-4). Mean accuracy (percentage) and standard deviation over three runs on a validation set (obtained sampling 20% of the images from the training set). Best results highlighted in bold.

Aggregation	Analytical form	CIFAR-10	CIFAR-100
Sum	$a_{\text{sum}}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i + \mathbf{z}_j$	57.60±0.23	29.45±0.69
Mean	$a_{\text{mean}}(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i + \mathbf{z}_j}{2}$	57.77±0.74	29.15±0.80
Maximum	$a_{\text{max}}(\mathbf{z}_i, \mathbf{z}_j) = \max(\mathbf{z}_i, \mathbf{z}_j)$	56.45±1.15	26.58±1.26
Concatenation	$a_{\text{cat}}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i \frown \mathbf{z}_j$	<b>60.81±0.25</b>	<b>32.36±0.73</b>

Table 14: Ablation of the relation head. The models have been trained on unlabeled CIFAR-10 and CIFAR-100 and tested on various benchmarks with a ResNet32 backbone for 200 epochs (mean accuracy and standard deviation of 3 runs). We consider three head types: (a) *dot product* between the pairs encoded through the backbone, followed by BCE loss; (b) *Encoder + dot product*, aggregation is not performed, for each encoded representation an MLP performs a second encoding, then dot product is applied between pairs and the BCE loss minimized (similar to SimCLR, Chen et al. 2020); (c) *Relation module* corresponds to the proposed method where encodings are aggregated (concatenation) and passed through an MLP for binary classification. All the other factors are kept constant for a fair comparison (e.g. augmentation strategy, mini-batch size). Best results in bold.

Head type	Linear Evaluation		Domain Transfer		CIFAR-100-20
	CIFAR-10	CIFAR-100	10→100	100→10	
(a) dot product	72.74±0.22	28.77±0.44	18.19±0.10	51.9±0.50	45.05±1.07
(b) Encoder + dot product	59.44±0.59	29.91±1.28	28.29±0.90	53.65±0.85	36.94±1.30
(c) Relation module (ours)	<b>74.99±0.07</b>	<b>46.17±0.17</b>	<b>41.50±0.35</b>	<b>67.81±0.42</b>	<b>52.44±0.47</b>

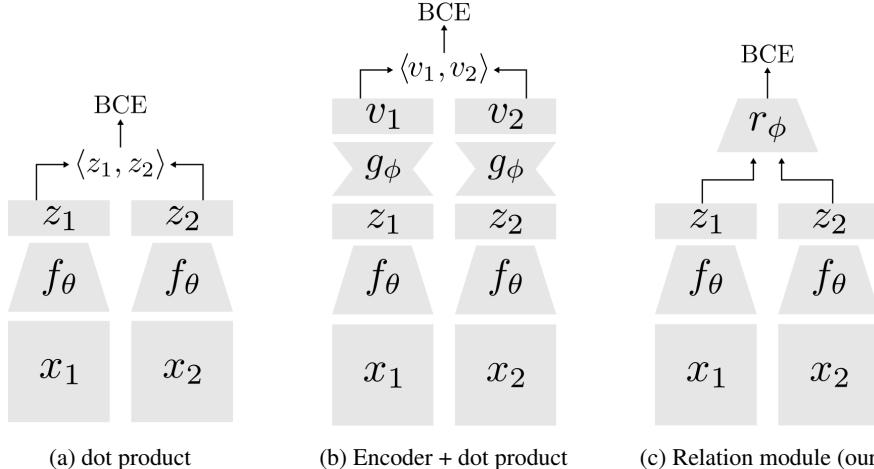


Figure 3: Ablation of the relation head (graphical illustration). Comparison between the two ablations in (a) and (b), and the full model in (c). In (a) the head is removed and the dot product  $\langle z_1, z_2 \rangle$  is used to compare the representations pair. In (b) the relation head is replaced with an encoder  $g_\phi$  that projects each representation in another latent space where the dot product is performed. In (c) is showed the full model, with the relation module  $r_\phi$  taking in input the aggregated pair. In all cases is minimized the binary cross-entropy loss (BCE) over positive and negative pairs.

## B.9 Image retrieval: qualitative analysis



Figure 4: Image retrieval given 25 random queries (not cherry-picked) on CIFAR-10 with ResNet-32. The query is the leftmost image (red frame), followed by the top-10 most similar images (Euclidean distance) in representation space. Comparison between (a) self-supervised relational reasoning (ours), and (b) self-supervised rotation prediction (Gidaris et al., 2018). Our method shows better distinction between categories which are hard to separate, (e.g. ships vs planes in row 4, trucks vs cars in row 12). Moreover, the lower sample variance and the higher similarity with the query, indicates a fine-grained organization in representation space (e.g. red sport cars in row 1, long white trucks in rows 11 and 12, deer with snow in row 16, blue car in row 22, dog breeds in row 25).

## B.10 Image retrieval: error analysis

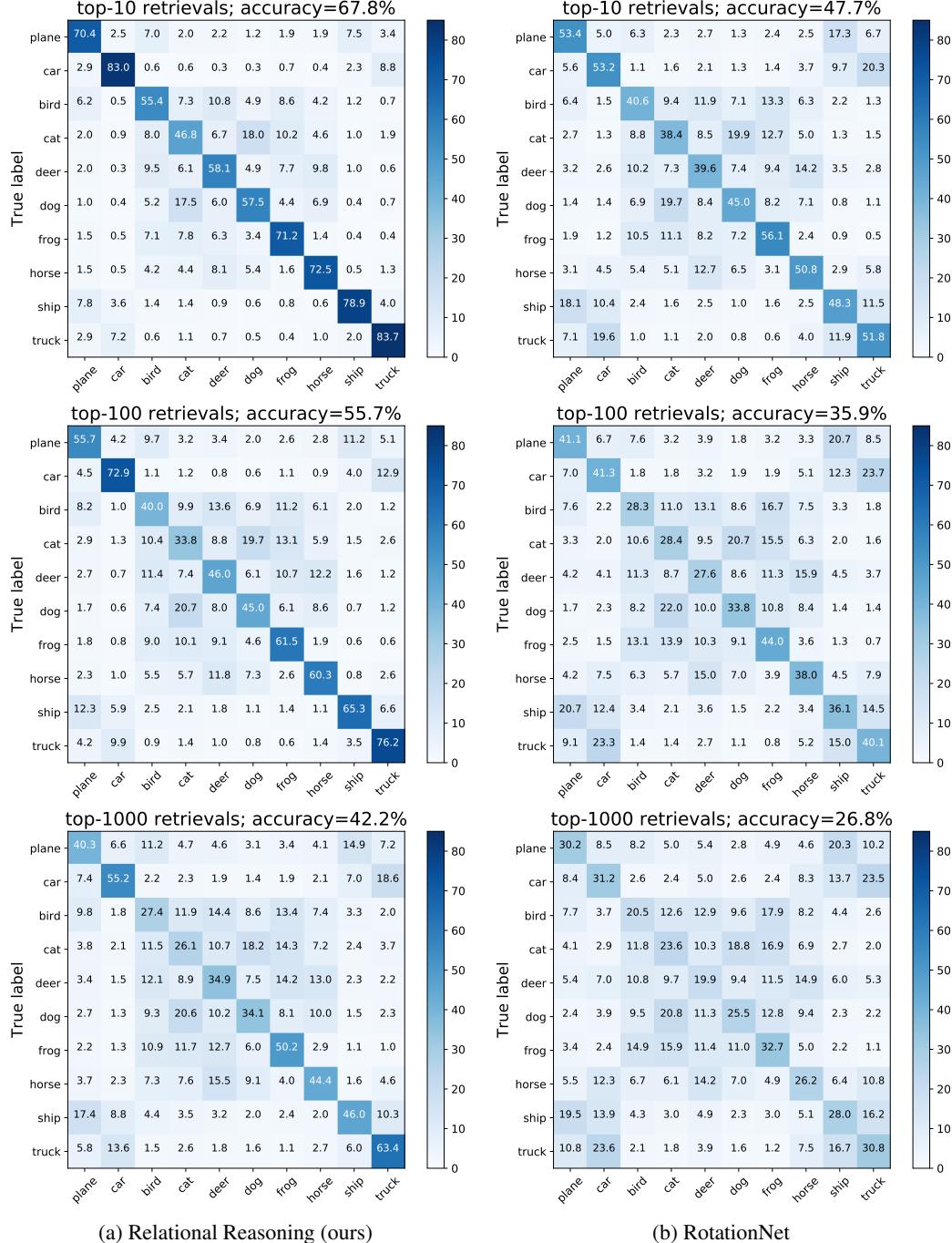


Figure 5: Confusion matrix obtained sampling 500 images per class (CIFAR-10, ResNet-32) and retrieving the top-10/100/1000 (top/middle/bottom table) closest images in representation space via Euclidean distance. Accuracy in percentage (three seeds) over correct retrievals (same category). Comparison between (a) self-supervised relational reasoning (ours), and (b) self-supervised rotation prediction (Gidaris et al., 2018). The proposed method shows a superior accuracy across all categories while being more robust against misclassification errors.

## B.11 Representations: qualitative analysis

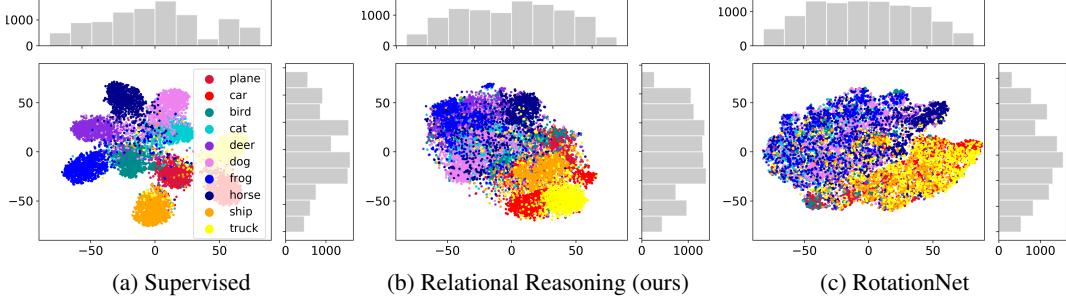


Figure 6: Visualization of t-SNE embeddings for the 10K test points in CIFAR-10. ResNet-32 backbone trained via (a) supervised learning, (b) self-supervised relational reasoning (ours), and (c) self-supervised rotation prediction (Gidaris et al., 2018). Our method shows a lower scattering, with clusters which are more distinct.

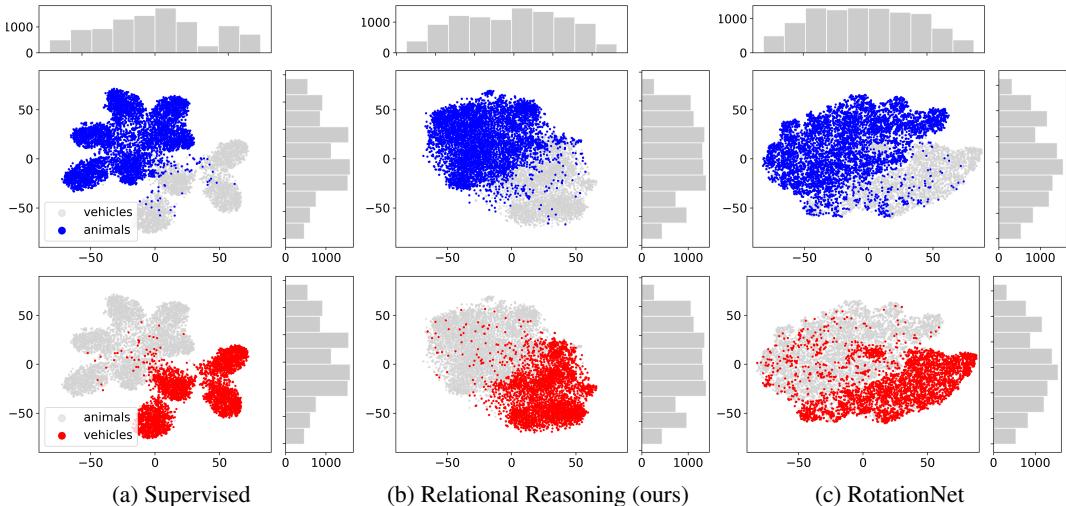


Figure 7: Visualization of t-SNE embeddings for the 10K test points in CIFAR-10 divided in two super-categories: vehicles (plane, car, ship, truck), and animals (bird, cat, deer, dog, frog, horse). ResNet-32 backbone trained via (a) supervised learning, (b) self-supervised relational reasoning (ours), and (c) self-supervised rotation prediction (Gidaris et al., 2018). Our method shows a better split, lower scattering, and a minor overlap between the two super-categories.

## C Pseudo-code of the method

---

**Algorithm 1** Self-supervised relational learning: training function and shuffling without collisions.

**Require:**  $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$  unlabeled training set;  $\mathcal{A}(\cdot)$  augmentation distribution;  $\theta$  parameters of  $f_\theta$  (neural network backbone);  $\phi$  parameters of  $r_\phi$  (relation module); aggregation function  $a(\cdot, \cdot)$ ;  $\alpha$  and  $\beta$  learning rate hyperparameters;  $K$  number of augmentations;  $M$  mini-batch size;

```

1: function TRAIN( $\mathcal{D}, \alpha, \beta, M, K, \theta, \phi$ )
2:   while not done do
3:      $\mathcal{B} = \{\mathbf{x}_m\}_{m=1}^M \sim \mathcal{D}$                                  $\triangleright$  Sampling a mini-batch
4:     for  $k = 1$  to  $K$  do
5:        $\mathcal{B}^{(k)} \sim \mathcal{A}(\mathcal{B})$                                  $\triangleright$  Sampling  $K$  mini-batch augmentations
6:        $\mathcal{Z}^{(k)} = f_\theta(\mathcal{B}^{(k)})$                                  $\triangleright$  Forward pass in the backbone
7:     end for
8:      $\mathcal{P} = \{\}$                                  $\triangleright$  Empty set to store aggregated pairs and targets
9:     for  $i = 1$  to  $K - 1$  do
10:      for  $j = i + 1$  to  $K$  do
11:         $\mathcal{P} \leftarrow (a(\mathcal{Z}^{(i)}, \mathcal{Z}^{(j)}), \mathbf{t} = 1)$      $\triangleright$  Aggregating and appending positive pairs
12:         $\tilde{\mathcal{Z}}^{(j)} = \text{SHUFFLE}(\mathcal{Z}^{(j)})$                  $\triangleright$  Shuffling without collisions
13:         $\mathcal{P} \leftarrow (a(\mathcal{Z}^{(i)}, \tilde{\mathcal{Z}}^{(j)}), \mathbf{t} = 0)$      $\triangleright$  Aggregating and appending negative pairs
14:      end for
15:    end for
16:     $\mathbf{y} = r_\phi(\mathcal{P})$                                  $\triangleright$  Forward pass in the relation module
17:     $\mathcal{L} = \text{BCE}(\mathbf{y}, \mathbf{t})$                                  $\triangleright$  Estimating the Binary Cross-Entropy loss
18:     $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$                  $\triangleright$  Updating backbone
19:     $\phi \leftarrow \phi - \beta \nabla_\phi \mathcal{L}$                  $\triangleright$  Updating relation module
20:  end while
21:  return  $\theta, \phi$                                  $\triangleright$  Returning the learned weights
22: end function

23: function SHUFFLE( $\mathcal{Z}$ )
24:    $\tilde{\mathcal{Z}} = \mathcal{Z}$                                  $\triangleright$  Copying the input set
25:   for  $m = 1$  to  $M$  do
26:      $\tilde{m} \sim \{1, \dots, M\} \setminus \{m\}$            $\triangleright$  Sampling an index  $\tilde{m} \neq m$ 
27:      $\tilde{\mathcal{Z}}_{\tilde{m}} \leftarrow \mathcal{Z}_{\tilde{m}}$                  $\triangleright$  Assigning a random representation with index  $\tilde{m}$ 
28:   end for
29:   return  $\tilde{\mathcal{Z}}$                                  $\triangleright$  Returning the shuffled set
30: end function

```

---

## D Essential PyTorch code of the method

### D.1 Data loader

```

import torchvision
from PIL import Image

class MultiCIFAR10(torchvision.datasets.CIFAR10):
    """Override torchvision CIFAR10 for multi-image management.
    Similar class can be defined for other datasets (e.g. CIFAR100).
    """
    def __init__(self, K, **kwds):
        super().__init__(**kwds)
        self.K = K # tot number of augmentations

    def __getitem__(self, index):
        img, target = self.data[index], self.targets[index]
        pic = Image.fromarray(img)

```

```

    img_list = list()
    if self.transform is not None:
        for _ in range(self.K):
            img_transformed = self.transform(pic.copy())
            img_list.append(img_transformed)
    else:
        img_list = img
    return img_list, target

```

## D.2 Augmentations

```

import torchvision.transforms as transforms

normalize = transforms.Normalize(mean=[0.491, 0.482, 0.447],
                                 std=[0.247, 0.243, 0.262]) # CIFAR10
color_jitter = transforms.ColorJitter(brightness=0.8, contrast=0.8,
                                       saturation=0.8, hue=0.2)
rnd_color_jitter = transforms.RandomApply([color_jitter], p=0.8)
rnd_gray = transforms.RandomGrayscale(p=0.2)
rnd_rcrop = transforms.RandomResizedCrop(size=32, scale=(0.08, 1.0),
                                           interpolation=2)
rnd_hflip = transforms.RandomHorizontalFlip(p=0.5)
train_transform = transforms.Compose([rnd_rcrop, rnd_hflip,
                                     rnd_color_jitter, rnd_gray,
                                     transforms.ToTensor(), normalize])

```

## D.3 Self-supervised relational reasoning

```

import torch

class RelationalReasoning(torch.nn.Module):
    def __init__(self, backbone):
        super(RelationalReasoning, self).__init__()
        feature_size = 64*2 # multiply by 2 since aggregation='cat'
        self.backbone = backbone
        self.relation_head = torch.nn.Sequential(
            nn.Linear(feature_size, 256),
            nn.BatchNorm1d(256),
            nn.LeakyReLU(),
            nn.Linear(256, 1))

    def aggregate(self, features, K):
        relation_pairs_list = list()
        targets_list = list()
        size = int(features.shape[0] / K)
        shifts_counter=1
        for index_1 in range(0, size*K, size):
            for index_2 in range(index_1+size, size*K, size):
                # Using the 'cat' aggregation function by default
                pos_pair = torch.cat([features[index_1:index_1+size],
                                      features[index_2:index_2+size]], 1)
                # Shuffle by rolling the mini-batch (negatives)
                neg_pair = torch.cat([
                    features[index_1:index_1+size],
                    torch.roll(features[index_2:index_2+size],
                               shifts=shifts_counter, dims=0)], 1)
                relation_pairs_list.append(pos_pair)
                relation_pairs_list.append(neg_pair)
                targets_list.append(torch.ones(size, dtype=torch.float32))
                targets_list.append(torch.zeros(size, dtype=torch.float32))
                shifts_counter+=1
            if(shifts_counter>=size):
                shifts_counter=1 # avoid identity pairs
        relation_pairs = torch.cat(relation_pairs_list, 0)

```

```

        targets = torch.cat(targets_list, 0)
        return relation_pairs, targets

    def train(self, tot_epochs, train_loader):
        optimizer = torch.optim.Adam([
            {'params': self.backbone.parameters()},
            {'params': self.relation_head.parameters()}])
        BCE = torch.nn.BCEWithLogitsLoss()
        self.backbone.train()
        self.relation_head.train()
        for epoch in range(tot_epochs):
            # the real target is discarded (unsupervised)
            for i, (data_augmented, _) in enumerate(train_loader):
                K = len(data_augmented) # tot augmentations
                x = torch.cat(data_augmented, 0)
                optimizer.zero_grad()
                # forward pass (backbone)
                features = self.backbone(x)
                # aggregation function
                relation_pairs, targets = self.aggregate(features, K)
                # forward pass (relation head)
                score = self.relation_head(relation_pairs).squeeze()
                # cross-entropy loss and backward
                loss = BCE(score, targets)
                loss.backward()
                optimizer.step()
                # estimate the accuracy
                predicted = torch.round(torch.sigmoid(score))
                correct = predicted.eq(targets.view_as(predicted)).sum()
                accuracy = (100.0 * correct / float(len(targets)))

            if(i%100==0):
                print('Epoch [{}][{}/{}] loss: {:.5f}; accuracy: {:.2f}%' \
                    .format(epoch+1, i+1, len(train_loader)+1,
                            loss.item(), accuracy.item()))

```

#### D.4 Main

```

def main():
    backbone = Conv4() # it should be a CNN with 64 linear output units
    model = RelationalReasoning(backbone)
    train_set = MultiCIFAR10(K=4, # it should be K=32 for CIFAR10/100
                             root='data', train=True,
                             transform=train_transform,
                             download=True)
    train_loader = torch.utils.data.DataLoader(train_set,
                                              batch_size=64,
                                              shuffle=True)
    model.train(tot_epochs=200, train_loader=train_loader)
    torch.save(model.backbone.state_dict(), './backbone.tar')

```