MEASURING THE EFFECTIVENESS OF SELF-SUPERVISED LEARNING USING CALIBRATED LEARNING CURVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Transfer learning has witnessed remarkable progress in recent years, for example, with the introduction of augmentation-based contrastive self-supervised learning methods. While a number of large-scale empirical studies on the transfer performance of such models have been conducted, there is not yet an agreed-upon set of control baselines, evaluation practices, and metrics to report, often hindering a nuanced and calibrated understanding of the real efficacy of the methods. We propose an evaluation standard that aims to quantify and communicate transfer learning performance in an informative and accessible setup. This is done by baking a number of simple yet critical control baselines in the evaluation method, particularly the ‘blind-guess’ (quantifying the dataset bias), ‘scratch-model’ (quantifying the architectural contribution), and ‘maximal-supervision’ (quantifying the upper-bound). To demonstrate how the proposed evaluation standard can be employed, we provide an example empirical study investigating a few basic questions about self-supervised learning. For example, using this standard, the study shows the effectiveness of existing self-supervised pre-training methods is skewed towards image classification tasks versus others, such as dense pixel-wise predictions.

1 INTRODUCTION

Creating computer vision models that can support a wide range of downstream tasks will require the development of representations whose utility extend beyond the exact same objective they were optimized for. Transfer learning is a general approach to operationalize this perspective by using representations of a model trained on a source task to improve the sample complexity of learning another downstream task. Supervised pre-training approaches to transfer learning, however, still have the drawback of requiring large annotated datasets for the source task. Self-supervised learning, in turn, holds the potential to remove this bottleneck by learning representations directly from unlabelled data by means of employing a proxy pre-training objective whose optimization results in transferable representations, but does not require manual labelling (Noroozi & Favaro, 2016; Zhang et al., 2016; Gidaris et al., 2018). For example, the recently revisited contrastive learning approaches (Hadsell et al., 2006; Chen et al., 2020a; He et al., 2020; Caron et al., 2020) constitute a milestone in this paradigm and were shown to achieve impressive transfer performance competitive with representations obtained in a fully supervised way for classification tasks.

Given the rapid progress in transfer and self-supervised learning, obtaining a clear and complete picture of the increasingly large number of methods being proposed is difficult, in part due to an absence of standard evaluation practices and the non-uniformity of reported (control) baselines. Ideally, such an evaluation standard should have a number of desirable properties, such as:

1. providing a simple metric that clearly communicates the “effectiveness” transfer learning of a given method on a natural and interpretable scale.
2. taking into account the effects of *irreducible factors that upper-bound the best possible transfer performance*, such as the uncertainty in observations or the imperfect nature of current neural network training practices.
3. normalizing away the benefits due to the statistical regularities of datasets that *lower-bound the worst possible performance* (Coughlan & Yuille, 2000), to prevent them from being a confounder in comparisons.
4. providing an intuitive way to draw comparisons about the utility of different self-supervised learning methods *across different downstream tasks*.

This paper proposes an evaluation standard that satisfies the aforementioned properties by incorporating a set of key control baselines in the evaluation. We then proceed with example analyses on questions that would be otherwise harder to study without such a standard – for example, assessing whether existing self-supervised learning methods are equally effective across different types of tasks. The proposed standard is not limited to self-supervised learning and is applicable to any transfer learning evaluation.

2 Related Work

Self-Supervised Representation Learning is a special case of general transfer learning and one of the most attended ones today. Earlier work on self-supervised learning largely focused on hand-designed proxy-tasks, such as solving a jigsaw puzzle (Noroozi & Favaro, 2016), image colorization (Zhang et al., 2016), rotation prediction (Gidaris et al., 2018), as well as numerous others (Misra & Maaten, 2020; Pathak et al., 2016; Zamir et al., 2016). More recently, a number of augmentation-based contrastive methods have been proposed (Chen et al., 2020a; He et al., 2020; Caron et al., 2020; Henaff, 2020; Chen et al., 2020b), and were shown to outperform previous approaches significantly. These methods employ a contrastive-loss formulation (Hadsell et al., 2006), and learn representations that are similar for augmented views of the same image and dissimilar for different images. Bardes et al. (2021); Zbontar et al. (2021); Grill et al. (2020) also suggest learning representations that are close for augmentations of the same image, but do not use negative examples explicitly and achieve similar performance to contrastive methods.

Benchmarking and Analysis. The literature contains various empirical studies on the effectiveness of self-supervised pre-training, across a wide range of methods and downstream tasks. Van Horn et al. (2021); Islam et al. (2021); Cole et al. (2021) focus on classification tasks and study how the final transfer performance depends on the image domain of a dataset, downstream-task complexity, and the available amount of labelled data for transfer. Ericsson et al. (2021); Newell & Deng (2020) further include non-classification tasks in their evaluation setup, and Kotar et al. (2021) conduct a comprehensive empirical study with a diverse set of downstream tasks to demonstrate the benefits of contrastive self-supervised pre-training. Goyal et al. (2019); Kolesnikov et al. (2019); Tian et al. (2020); Xiao et al. (2020); Cole et al. (2021) examine the effects of various pre-training parameters such as the model capacity, pretext task complexity, and the augmentation policy used in the contrastive loss formulation, to provide a further understanding of the successes of self-supervised pre-training. Tian et al. (2020); Tsai et al. (2020); Arora et al. (2019); Lee et al. (2020) build a mathematical framework to explain how self-supervised pre-training can improve the performance on downstream tasks, aiming to provide theoretical guarantees. In this work, we also include non-classification tasks in our showcase analysis, and aim to study the transfer performance more comprehensively with the proposed control baselines and evaluations across different data-regimes.

Evaluation Metrics and Baselines. A standard way to evaluate self-supervised learning methods is through freezing their pre-trained representations and reporting the performance of a linear classifier trained on top to solve ImageNet classification. Ericsson et al. (2021) show empirically that this practice is not as predictive of the downstream performance for non-classification tasks as it is for classification tasks. Kotar et al. (2021) adopt a more comprehensive setup that evaluates transfer performance on a broader range of downstream tasks, and opt for reporting most of their results in terms of relative improvements over ImageNet supervised pre-training, but the absolute level of performance of the ImageNet pre-training baseline remains unclear for their study. Newell & Deng (2020) and Cole et al. (2021) include comparisons to a model trained from scratch on the same amount of labelled data, and the former further suggests measuring the transfer efficacy as the proportion of labelled data saved by self-supervised pre-training compared to training a model from scratch to achieve the same performance. Our work also incorporates the scratch performance as a baseline to account for the difficulty of the downstream task. We additionally put our comparisons into perspective by attempting to account for the dataset bias and the performance upper-bound due to irreducible uncertainties or architectural limitations.

3 The Evaluation Standard

This section describes the control baselines and the overall evaluation standard we propose. We start with a short description of the transfer-learning setting, and then proceed with presenting the definitions and motivations behind each of the control baselines. We then incorporate these baselines

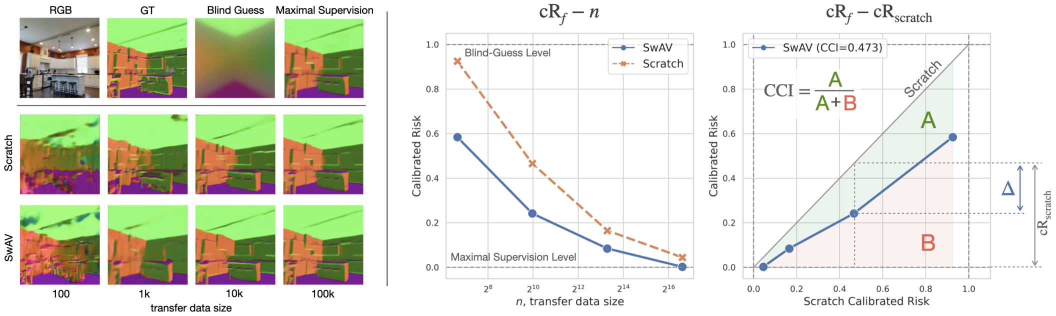

Figure 1: Right: Exemplary $\mathbf{cR}_{\boldsymbol{f}}\text{-}\boldsymbol{n}$ and $\mathbf{cR}_{\boldsymbol{f}}\text{-}\mathbf{cR}_{\mathbf{scratch}}$ curves for the SwAV self-supervised pre-training method, transferred to surface normals estimation. The relative improvement against the scratch control baseline (i.e. $\Delta/\mathrm{cR}_{\mathrm{scratch}}$) can directly be read from the plot and provides a visualization for the transfer efficacy of a given method. Left: corresponding qualitative surface normals predictions for SwAV and three control baselines: blind-guess, maximal-supervision and scratch. In this example, the improvement of the SSL pre-training over scratch is observed to diminishes with more training data, which can also be inferred from the qualitative examples for 10K/100K. This (rather general) trend indicates the benefit of transfer learning in high-data regime is less significant, and therefore, relative improvement comparisons are more meaningful across different data-regimes (particularly mid and low data) where the scratch baseline performs relatively poorly.

into an affine rescaling of the empirical risk, and discuss the associated visualizations and additional metrics.

3.1 Transfer Learning Setting

We consider a standard transfer learning setting that employs an encoder-decoder architecture. The *encoder* $\psi : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{X}$ and $\mathcal{Z}$ denote the set of input images and their projected latent representations respectively, is first optimized using a given *pre-training method* (e.g., a self-supervised method, or a source-task solved in a fully supervised way) on a *pre-training dataset* $\mathcal{D}_{\mathrm{pre}}$. The *decoder* $\phi : \mathcal{Z} \to \mathcal{Y}$ for the *downstream task*, with $\mathcal{Y}$ denoting the output space, is then trained in a supervised way using a *transfer dataset* $\mathcal{D}$ of annotated images to minimize the task-specific loss $\mathcal{L}$, and the encoder parameters are also fine-tuned. The number of images $|\mathcal{D}|$ used for this optimization is called the *data-regime* for transfer learning. The resulting model is denoted as $f_\theta = \phi \circ \psi$, where θ includes all the network parameters. After training, the empirical risk is computed over the test partition $\mathcal{D}_{\mathrm{test}}$ of the transfer dataset using the loss $\mathcal{L}$ for the downstream-task:

$$R_{f_\theta} = \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{(x,y)\in\mathcal{D}_{\mathrm{test}}} \mathcal{L}(y, f_\theta(x))). \tag{1}$$

3.2 Control Baselines

Motivation. Transfer learning results are commonly reported using the empirical risk. Therefore, interpreting such results requires knowledge about the scale of the training loss and its nonlinear relation to the inherent prediction fidelity/value for the particular downstream task and dataset. To illustrate with a toy example, when a classification model is reported to achieve the accuracy of 0.9, judging whether this model performs well or not cannot be done solely by means of the reported metric: the model can be considered proficient if there are 1000 uniformly distributed classes in the dataset, but not so if 90% of the images belong to the same class – since the accuracy of merely statistically informed guess is 0.001 ($900\times$ worse than the model's) for the former and 0.9 (the same as the model's) for the latter. Furthermore, the sample complexity of obtaining an *additional improvement* will differ depending on the granularity and similarity of the categories.

Such observations motivate the following questions: What additional context can we provide to create a more complete picture of the transfer performance? What are the most sensible, efficient, and simple control baselines that would allow us to attain a more accurate assessment?

The Scratch Control Baseline: When training a decoder ϕ on a transfer dataset, choices about training parameters (e.g., the model architecture or the optimization method used) are an important source of inductive bias. Similarly, those specifics of a downstream task and dataset that characterize their amenability for model training have an influence on how the optimization will proceed. Such considerations are independent of the particular choice of pre-training method being evaluated, yet they have a critical impact on the final transfer performance. Therefore, as previous studies also illustrate (Cole et al., 2021), it is important to include the *scratch control baseline* (i.e., a randomly initialized model trained only using a transfer dataset corresponding to the data-regime being examined), to disentangle and clarify the benefits of the pre-training method.

The Maximal-Supervision Control Baseline: Even in the presence of an infinite amount of data, many factors can still limit the minimum error achievable on a downstream task. Examples include uncertainties and multi-modality in the label generation process, finite model capacities and architectural bottlenecks, or the non-convexity of the optimization problem and the imperfect nature of the algorithms employed. Having an estimate of such factors is useful for framing the transfer performance in an interpretable scale. We, therefore, employ a *maximal-supervision control baseline* – a randomly initialized model trained using a large amount of data for the downstream task intended to provide *an approximation* to the maximum achievable performance on the downstream task. This baseline may not always be attainable, e.g., when sufficiently large training data for the dataset domain or a theoretical approximation are unavailable. In our experiments, we will show cases where a good approximation is attainable (e.g., for Tasknomomy (Zamir et al., 2018)) as well as cases where it is not attainable (e.g., for CIFAR-100 (Krizhevsky et al., 2009)), and discuss the implications.

The Blind-Guess Control Baseline: Regularities in datasets can allow statistically informed guesses for the downstream task (Coughlan & Yuille, 2000). Such guesses can often be unexpectedly performant (Zamir et al., 2018; Eftekhar et al., 2021). The effects of such regularities can be accounted for by constructing an input-agnostic *blind-guess control baseline*, defined as the single constant prediction with the lowest error $\arg\min_{\hat{y}\in\mathcal{Y}} \mathbb{E}_y\left[\mathcal{L}(\hat{y}, y)\right]$, which corresponds to the mean output for L_2 loss, the median output for L_1 loss, and the most-represented class for the 0-1 loss. This control baseline serves as a sanity check to clarify whether a pre-training method provides a tangible benefit, as compared to solely capturing regularities in dataset statistics in an input-independent prediction.

Importance of Control Baselines: The main goal of incorporating control baselines is to project the performance of a method onto a comparative scale to capture *how well it works*, as well as to provide insight into *why it works* through rejecting a set of null-hypotheses. Computer vision literature contains numerous recent examples of how the inclusion of appropriate baselines can provide such a clarification for their respective fields and problems. For example, Tatarchenko et al. (2019) employ a set of recognition baselines for the single-view 3D reconstruction problem, and reveals that the performance of state of the art models are statistically indistinguishable from classification and retrieval based methods. Another example is the study by Zamir et al. (2018), which employ a gain metric (i.e., the win rate against a network trained from scratch) to show that given a target task (e.g., classification), multiple source tasks (e.g., colorization) can collapse to a performance level below the scratch baseline in the lower data regime. Sax et al. (2019) employ a *blind* intelligent actor baseline (i.e., a policy operating without visual input) for navigation tasks, and reveals that policies trained from scratch as well as state of the art representation learning methods perform at a similar level to this baseline when tested on unseen environments. Here we use such baselines in a more systematic way.

Transfer learning is a field where empirical studies are prevalent, and this empirical nature makes rendering evaluations with respect to an appropriate set of baselines particularly important for measuring the progress in a more accessible and comparable way, while also systematically revealing and preventing blind spots. The current common practice is to report and compare downstream performances between different pre-training methods using the original loss scale or relative to the ImageNet features performance (Kotar et al., 2021; Ericsson et al., 2021; Islam et al., 2021). While this setup still allows conclusions based on quantitative results, it is hard to assess whether the difference in the methods' performance is significant or not. As an example, comparing the L_1 loss for the downstream task of depth estimation as reported in Fig.7 of (Kotar et al., 2021) with the blind baseline reported by (Zamir et al., 2020) in Tab. 1, one can conclude that all models perform at the blind guess level, and no meaningful improvement was made. Employing an appropriate set of common control baselines would make it easier to identify and prevent miscalibrations across different studies and gain a more complete and comparable picture about the real effectiveness of methods.

3.3 The Proposed Evaluation Setup

Calibrated-Risk: To incorporate all three of the previously proposed control baselines in a single metric, we calibrate the empirical risk R_f of a transfer model f as follows:

$$\mathrm{cR}_f = \frac{R_f - R_{\max}}{R_{\mathrm{blind}} - R_{\max}}, \tag{2}$$

where R_{blind} and $R_{\max}$ correspond to the empirical risks of the blind-guess and maximal-supervision controls. We refer to cR_f as the *calibrated risk*. Similarly, the calibrated risk of the scratch control is computed as:

$$\mathrm{cR}_{\mathrm{scratch}} = \frac{R_{\mathrm{scratch}} - R_{\max}}{R_{\mathrm{blind}} - R_{\max}}, \tag{3}$$

where $R_{scratch}$ denotes the empirical risk of the scratch control baseline. It is useful to note that cR_f and $\mathrm{cR}_{scratch}$ are invariant to affine transformations of the training loss $\mathcal{L}$.

The calibrated risk of a given pre-training method varies depending on the *data-regime*, and comparisons across different pre-training methods can therefore show rank reversals. Accordingly, considering the performance *across a range of different data regimes*, in the form of a learning curve (Hoiem et al., 2021), can provide a more revealing picture of the rigidity and practical value of inductive biases introduced by different pre-training methods. As illustrated in Fig.1, we suggest two plots to illustrate the resulting curves:

- Plotting cR_f and $\mathrm{cR}_{\mathrm{scratch}}$ on the y-axis against the transfer dataset size n on the x-axis (i.e., $\mathbf{cR}_{\boldsymbol{f}}\text{-}\boldsymbol{n}$ curves, as shown in Fig.1, middle).
- Plotting cR_f on the y-axis against $\mathrm{cR}_{\mathrm{scratch}}$ on the x-axis, to emphasize the relative improvement against scratch performance (i.e., $\mathbf{cR}_{\boldsymbol{f}}\text{-}\mathbf{cR}_{\mathbf{scratch}}$ curves, as shown in Fig.1, right).

Reading $\mathbf{cR}_{\boldsymbol{f}}\text{-}\boldsymbol{n}$ Curves: Eq. 2 implies that calibrated risks for the maximal-supervision and blind-guess control baselines correspond to horizontal lines $\mathrm{cR}_f = 0$ and $\mathrm{cR}_f = 1$. If the maximal-supervision control baseline is trained using sufficiently many data samples (i.e., meaning that $R_{\max}$ is a good approximation of the maximum achievable performance, which may not always be true for small-scale datasets), then $R_f \leq R_{\max}$ holds for all data regimes and per Eq. 2, cR_f approaches the maximal-supervision level $\mathrm{cR}_f = 0$ as n goes to infinity. The cR_f curve of a good pre-training method should lie below the $\mathrm{cR}_{\mathrm{scratch}}$ curve and as close to the line $\mathrm{cR}_f = 0$ as possible across all data regimes.

Reading $\mathbf{cR}_{\boldsymbol{f}}\text{-}\mathbf{cR}_{\mathbf{scratch}}$ Curves: In most situations, the absolute scalar value cR_f by itself does not provide a clear indication of how well a pre-training method performs, and the more informative quantity is instead the *relative improvement* with respect to the scratch performance $\mathrm{cR}_{\mathrm{scratch}}$. To visually facilitate this comparison, we propose plotting cR_f against $\mathrm{cR}_{\mathrm{scratch}}$. This maps the $\mathrm{cR}_{\mathrm{scratch}}$-$n$ curve onto the main diagonal $x = y$, and a good pre-training method should therefore lie below it. The relative-improvement of a pre-training method over scratch can be read from the ratio of distances Δ and $\mathrm{cR}_{\mathrm{scratch}}$ in Fig. 1.

The Choice of Transfer Data-Regimes. A qualitative assessment of the images given in Fig. 1 and Fig. 3 illustrates how the efficiency of self-supervised pre-training compared to the scratch baseline depends on the amount of data available for transfer. Both methods perform similarly in high data-regimes and approach the performance of the maximal-supervision control, which can also be seen from qualitative examples for 100k train sizes. It is important, therefore, to choose training sizes appropriately when evaluating the efficiency of SSL pre-training. We define low- and high-data regimes as ones where the scratch performance is close to blind-guess and maximal-supervision controls respectively, and suggest choosing training-dataset sizes to cover this range.

Calibrated Cumulative Improvement. In order to facilitate tabular analysis and allow straightforward comparisons between different pre-training methods, we introduce a global *Calibrated Cumulative Improvement* (CCI) metric. We calculate it as the area between an SSL curve and the main diagonal as measured on the $\mathbf{cR}_{\boldsymbol{f}}\text{-}\mathbf{cR}_{\mathbf{scratch}}$ plot, divided by the total area under the main diagonal, shown in Fig. 1. CCI shows the total improvement of a pre-training method compared to scratch over multiple data-regimes. The CCI values of our experiments are available in the legends of each figure.

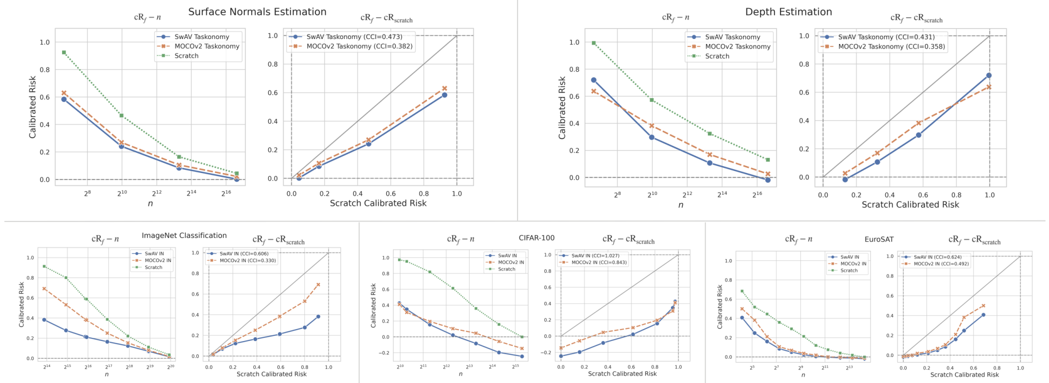

Figure 2: *By how much does contrastive self-supervised learning outperform training from scratch?* We plot $\mathbf{cR}_{\boldsymbol{f}}$-$\boldsymbol{n}$ and $\mathbf{cR}_{\boldsymbol{f}}$-$\mathbf{cR}_{\mathbf{scratch}}$ curves for SwAV and MoCov2 across different downstream tasks. For the three classification tasks, we use encoders pretrained on ImageNet (denoted as IN). For surface normals and depth estimation, we use the Taskonomy pre-trained versions of the same backbones. For all the tasks, contrastive SSL performs better than no pre-training by a relatively large margin, and SwAV outperforms MoCov2 in most cases.

4 Experiment Settings

To illustrate how the proposed control-baselines and evaluation standard can be employed, we perform an example empirical analysis of transfer learning in Section 4. The current section describes our experimental setup including pre-training methods, datasets, and downstream-tasks. All experiments were conducted using PyTorch (Paszke et al., 2017), and the associated code will be made available online.

4.1 Datasets and Downstream Tasks

The literature on benchmarking transfer performances of different pre-training methods commonly focuses on *image classification tasks* (Cole et al., 2021; Kolesnikov et al., 2020; Islam et al., 2021). Our experiments similarly incorporate three examples of such classification problems, namely ImageNet (Deng et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and EuroSAT (Helber et al., 2019). In addition to classification, we also include *pixel-wise regression tasks* using the Taskonomy dataset (Zamir et al., 2018). In total, we use four different datasets, each representing a distinct image domain with different properties.

Semantic Classification Tasks. ImageNet is a standard large-scale computer vision dataset of 1.3M natural images, each containing an object from one of the 1000 classes. We also use the relatively smaller CIFAR-100 and EuroSAT datasets to evaluate how well self-supervised learned representations transfer to different domains when only a small amount of labelled data is available. For ImageNet and CIFAR-100, we use the standard train/test splits, and split the EuroSAT dataset randomly. We train all models using the cross-entropy loss and evaluate them on the corresponding test sets using error rate. As the blind-guess prediction, we use the most common class. We train the maximal-supervision control baselines for CIFAR-100 and EuroSAT from scratch, using all available training images and use weights available from the PyTorch library (Paszke et al., 2017) for ImageNet.

Pixel-Wise Regression Tasks. We use two common pixel-wise regression tasks: depth and surface normals estimation from the Taskonomy dataset (Zamir et al., 2018). Taskonomy contains 4M images of natural indoor scenes from about 600 different buildings. We use images from the official full+ split buildings, and fix a random subset of 60K images from the set of test buildings to evaluate the final performances. We use L_1-norm as the loss for training on both downstream tasks. The blind-guess control baseline for both tasks is computed as the pixel-wise median over a set of 60K train images. The maximal-supervision control baselines are trained from scratch using 1.2M images from the same full+ split, and reach a level of performance close to the one reported in (Zamir et al., 2020).

4.2 Pre-Training Methods

Contrastive Self-Supervised Learning. In our showcase analysis, we consider representative samples of contrastive self-supervised learning methods: SwAV (Caron et al., 2020), MoCov2 (Chen et al., 2020c), SimCLR (Chen et al., 2020a), SimSiam (Chen & He, 2021), Barlow Twins (Zbontar et al., 2021), and PIRL (Misra & Maaten, 2020). These methods have the main attributes of modern contrastive methods and most of them achieve state-of-the-art performance on standard benchmarks. We use the pre-trained models from the corresponding github releases[1], as well as pre-trained models from the VISSL library [2] (Goyal et al., 2021).

Non-Contrastive Pretext Tasks. As examples of non-contrastive approaches, we consider two proxy tasks based on image colorization (Zhang et al., 2016) and jigsaw puzzle (Noroozi & Favaro, 2016). For both proxy tasks, we use encoders pre-trained on ImageNet from the VISSL repository (Goyal et al., 2021).

Supervised Pre-Training. We consider two supervised tasks in our study to measure whether there is a gap between supervised and self-supervised pre-training approaches. First, we use a standard ImageNet pre-trained encoder, as commonly employed by many evaluation studies, and generally adopted as an initialization in the community. Second, for depth and surface normals estimation, we choose the corresponding pre-training tasks from the Taskonomy dictionary that were empirically found to result in the best transfer performance. As reported by Zamir et al. (2018), this task is reshading for both. We pre-train the reshading encoder on the Taskonomy dataset using the same number of images as ImageNet pre-training. For the ImageNet encoder we use the weights provided by PyTorch.

4.3 Architecture and Training Details

All pre-trained encoders share the same ResNet-50 architecture (He et al., 2016). For transfers to downstream tasks, we use two types of decoders. For classification tasks, we use a single fully-connected layer that takes the output of the final encoder's layer and outputs the logits for each class. For pixel-wise regression tasks, we use a UNet-style (Ronneberger et al., 2015) decoder with six upsampling blocks and skip-connections from the encoder layers of the same spatial resolution. We randomly sub-sample n images from the corresponding training set to form each data-regime and split them into train and validation sets. We use the validation split for early-stopping, as well as to choose the hyper-parameters for different downstream tasks. The final performance is evaluated on the fixed test split.

5 Empirical Analysis

The main goal of this section is to showcase an actual application of the proposed evaluation standard on a set of exemplary scenarios. It presents a targeted and small-scale empirical study that poses a set of questions about the transfer performance of self-supervised pre-training methods, and proceeds with a subsequent analysis that utilizes the suggested controls and visualizations to address them.

5.1 By how much does contrastive self-supervised learning outperform training from scratch?

To address this question, we visualize the $\mathbf{cR}_{\boldsymbol{f}}$-$\mathbf{cR}_{\mathbf{scratch}}$ and $\mathbf{cR}_{\boldsymbol{f}}$-$\boldsymbol{n}$ curves for MoCov2 and SwAV pre-training methods. For both methods, we use Taskonomy pre-trained encoders for pixel-wise regression tasks and ImageNet pre-trained versions for classification tasks, as we found them to perform better. The resulting curves are shown in Fig. 2.

We observe that both contrastive self-supervised methods outperform training from scratch on all downstream tasks in all data-regimes with a relatively large margin which, as one would expect, diminishes with more labelled data available for transfer. Between two pre-training methods, SwAV outperforms MoCov2 in most cases.

[1] SwAV: `https://github.com/facebookresearch/swav`, MoCov2: `https://github.com/facebookresearch/moco/`

[2] VISSL: `https://github.com/facebookresearch/vissl`

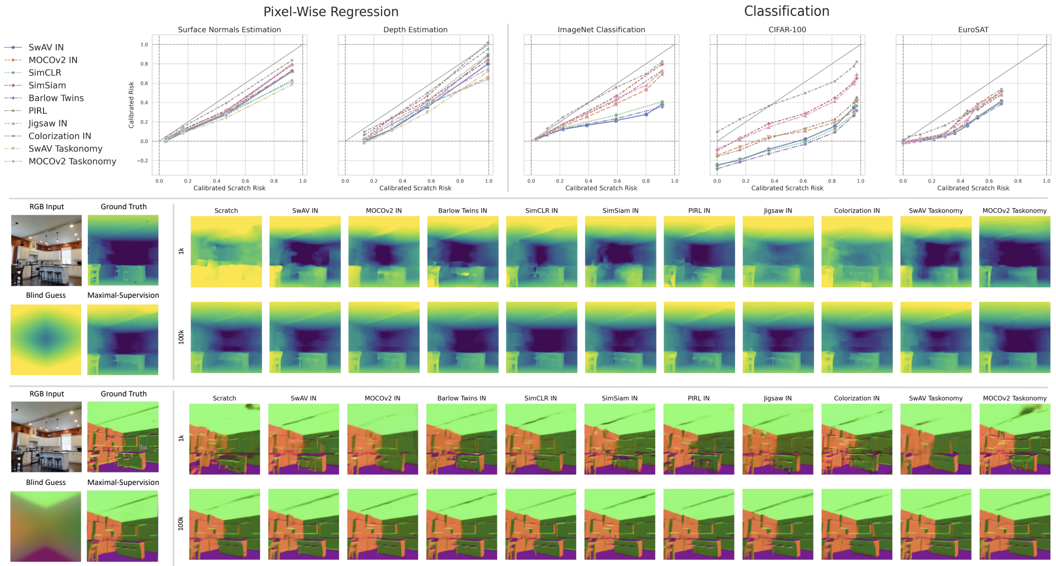

Figure 3: *Do different tasks benefit differently from self-supervised pre-training?* The $\mathbf{cR}_{\boldsymbol{f}}\text{-}\mathbf{cR}_{\mathbf{scratch}}$ plots above give a comparison across all pre-training methods for each downstream task. The images given below show corresponding visualizations of depth and normals predictions for different methods at two different dataset sizes. It can be observed that I. the differences among different pre-training methods are more pronounced for classification tasks compared to pixel-wise regression tasks, and II. the best pre-training result for classification is notably better than that of dense regression. Particularly at the high data-regime, such differences become insignificant. *This suggests the development of pre-training methods may be more curated towards downstream classification tasks.*

We note that the curves for CIFAR-100 eventually fall below the x axis. These negative calibrated risk values mean that transfer learning eventually starts outperforming the maximal-supervision control baseline (i.e., training from scratch using a large amount of available data) due to the small scale of the CIFAR-100 dataset. *The relative improvement and CCI computations with respect to the scratch performance, as well as comparisons between different methods for the same task are still valid in this case*, and the main difference is that the curves no longer converge to the maximal-supervision baseline since it is no longer a good approximation of the best achievable performance on the downstream task.

5.2 Do different tasks benefit differently from self-supervised pre-training?

The plots in Fig. 3 show how $\mathbf{cR}_{\boldsymbol{f}}\text{-}\mathbf{cR}_{\mathbf{scratch}}$ curves for different pre-training methods compare on five different downstream tasks. We observe that the transfer performance differences between the pre-training methods are larger on classification tasks and much less pronounced on pixel-wise regression ones. Additionally, in the high data regime the differences between the methods become insignificant for pixel-wise regression tasks, which can also be qualitatively confirmed from the visualizations of depth and normals predictions at two different dataset sizes. These observations suggest that *the development of pre-training methods may be implicitly biased towards classification tasks, rather than rather general purpose representations*.

5.3 By how much do contrastive methods outperform non-contrastive methods?

Recent works demonstrate that for semantic classification tasks, contrastive pre-training approaches result in better transfer performance compared to non-contrastive pretext tasks (He et al., 2020; Chen et al., 2020a). In this section, we investigate if the same trend holds for pixel-wise regression tasks. In Fig. 4, we show transfer performance results for colorization and jigsaw as compared to contrastive methods. For classification tasks, we replicate the common observation that contrastive methods provide a significant improvement compared to non-contrastive methods. However, we again observe that the difference is less pronounced on pixel-wise regression tasks, echoing the conclusion of Section 5.2.

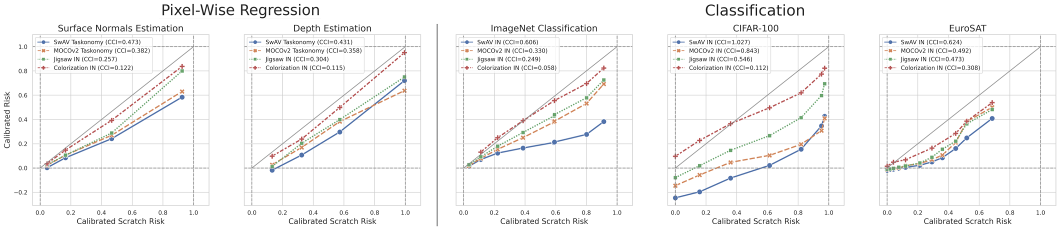

Figure 4: *By how much do contrastive methods outperform non-contrastive methods?* We plot the $\mathbf{cR}_{\boldsymbol{f}}\textbf{-}\mathbf{cR}_{\mathbf{scratch}}$ curves for contrastive pre-training methods (SwAV and MoCov2) together with non-contrastive methods (colorization and jigsaw). Colorization pre-training falls largely behind all other methods. Jigsaw pre-training, however, performs relatively similar to the two contrastive methods on pixel-wise regression tasks and EuroSAT classification, while the gap on CIFAR-100 and ImageNet is much larger (see Sec.5.2).

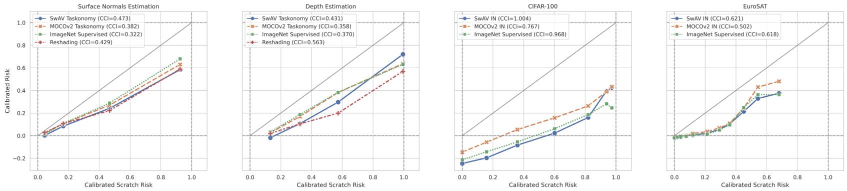

Figure 5: *Does self-supervised Learning outperform supervised pre-training?* We plot the $\mathbf{cR}_{\boldsymbol{f}}\textbf{-}\mathbf{cR}_{\mathbf{scratch}}$ curves for supervised and contrastive pre-training methods. ImageNet pre-training as a supervised method is included in the plots for all downstream tasks. For depth and normals estimation, we also include an encoder pre-trained on the reshading task, which is the optimal supevised transfer domain as reported by (Zamir et al., 2018). *We observe that supervised pre-training performs similarly to contrastive methods in most cases*, except reshading supervised pre-training on depth estimation in low-data regimes where it outperforms self-supervised counterparts (discussion in Sec.5.4).

5.4 Does self-supervised pre-training outperform supervised pre-training?

ImageNet Supervised Pre-Training. In Fig. 5, we compare ImageNet pre-training and contrastive methods. We observe that contrastive methods generally perform comparable or better than ImageNet pre-training. However, in terms of relative improvement against the scratch control baseline, this difference was found to be relatively minor.

Best Informed Pre-Training. For depth and normals estimation, we include further comparisons against the best pre-training task according to Taskonomy (Zamir et al., 2018) (see Section 4.2 for more details). The transfer results are presented in Fig. 5. We did not observe a large difference between contrastive and supervised pre-training for normals prediction, while the difference is more pronounced for depth estimation. This seems consistent with (Zamir et al., 2018) where the best source tasks were found to work well *if* they it is a close match with the target.

6 Conclusion and Limitations

We put forth an evaluation standard for measuring the effectiveness of pre-training methods for transfer learning, which incorporates three important control baselines:

- the *scratch* control baseline to disentangle the benefits of transfer learning from *the inductive biases introduced by other training choices* that are independent from the particular pre-training method,
- the *maximal-supervision* control baseline to account for *the effects of irreducible factors inherent to the downstream-task* that upper-bound the best possible transfer performance,
- the *blind-guess* control baseline to normalize away the benefits of *the statistical regularities of particular datasets* that lower-bound the worst possible transfer performance.

In Section 5, we used these control baselines to define the *calibrated risk* metric cR_f, with the goal of framing downstream task dependent empirical risk in an interpretable scale. We further showed two visualizations to plot calibrated risk curves, with the overall goal of providing an accessible way to judge the transfer efficacy of a given pre-training method, as well as to facilitate comparisons across methods. In Section 4, we have provided a targeted small-scale empirical study that showcases how the proposed evaluation standard can be applied. Through this study, we mainly observed that:

- Classification tasks consistently benefit more from contrastive self-supervised pre-training, compared to non-classification tasks such as pixel-wise regression (depth and surface normals estimation).
- The differences between self-supervised pre-training methods (and particularly between contrastive and non-contrastive approaches) are much less pronounced for non-classification tasks, as compared to classification tasks where there is a clear improvement from using contrastive methods. These are useful indications and insights for self-supervised learning research toward developing truly more general representation.

Our study, however, comes with several limitations:
Cross-Task Comparisons. Comparing methods across different tasks using their calibrated risk values makes the assumption that the prediction fidelity of each task can be captured by an affine transformation of its task-specific loss function, which may not hold true in practice.
Other Benefits of SSL. Besides reducing the need to label data, self-supervised learning provides other benefits, such as enabling training on continuous data streams (as opposed to fixed training sets) or reducing the reliance on rigid category definitions. This paper focused on quantifying the effectiveness of self-supervised learning in terms of transfer learning and reducing labeled data demands only.
Pre-training and Downstream Tasks Diversity. In comparison to other works whose main focus is to provide a comprehensive empirical study (Kotar et al., 2021; Chaves et al., 2021; Goyal et al., 2019; Islam et al., 2021; Van Horn et al., 2021), this paper mainly aims to propose an evaluation standard, and consequently we consider a more limited set of pre-training and downstream-tasks mainly curated to showcase the importance of the proposed control baselines.
The Effects of Pre-training Dataset Size. We evaluate the transfer efficacy as a function of the transfer dataset size, but the dependency on the pre-training dataset size is another important factor that determines the overall transfer performance, as studied in (Goyal et al., 2019; He et al., 2020).

References

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.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Levy Chaves, Alceu Bissoto, Eduardo Valle, and Sandra Avila. An evaluation of self-supervised pre-training for skin-lesion analysis. *arXiv preprint arXiv:2106.09229*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

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

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.

James Coughlan and Alan L Yuille. The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. *Advances in Neural Information Processing Systems*, 13:845–851, 2000.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10786–10796, 2021.

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*, pp. 5414–5423, 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

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*, pp. 6391–6400, 2019.

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.

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. *arXiv preprint arXiv:2006.07733*, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.

Derek Hoiem, Tanmay Gupta, Zhizhong Li, and Michal Shlapentokh-Rothman. Learning curves for analysis of deep networks. In *International Conference on Machine Learning*, pp. 4287–4296. PMLR, 2021.

Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. *arXiv preprint arXiv:2103.13517*, 2021.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1920–1929, 2019.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Klemen Kotar, Gabriel Ilharco, Ludwig Schmidt, Kiana Ehsani, and Roozbeh Mottaghi. Contrasting contrastive self-supervised representation learning pipelines. In *ICCV*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7345–7354, 2020.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

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*, pp. 2536–2544, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Alexander Sax, Jeffrey O. Zhang, Bradley Emi, Amir Zamir, Silvio Savarese, Leonidas Guibas, and Jitendra Malik. Learning to navigate using mid-level visual priors, 2019.

Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3405–3414, 2019.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

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*, pp. 12884–12893, 2021.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.

Amir R Zamir, Tilman Wekel, Pulkit Agrawal, Colin Wei, Jitendra Malik, and Silvio Savarese. Generic 3d representation via pose estimation and matching. In *European Conference on Computer Vision*, pp. 535–553. Springer, 2016.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11197–11206, 2020.

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.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.