Know Your Self-supervised Learning: A Survey on Image-based Discriminative Training

Anonymous authors
Paper under double-blind review

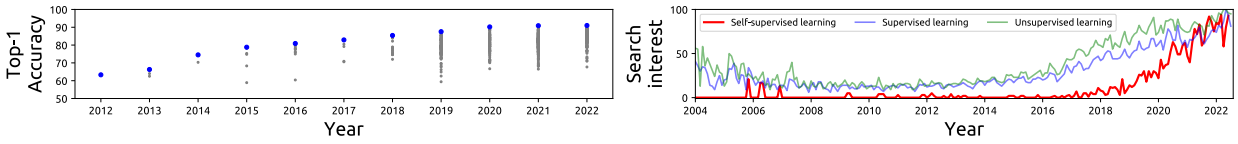
Abstract

Although supervised learning has been highly successful in improving the state-of-the-art in the domain of image-based computer vision in the past, the margin of improvement has diminished significantly in recent years, indicating that a plateau is in sight. Meanwhile, the use of self-supervised learning (SSL) for the purpose of natural language processing (NLP) has seen tremendous successes during the past couple of years, with this new learning paradigm yielding powerful language models. Inspired by the excellent results obtained in the field of NLP, self-supervised methods that rely on clustering, contrastive learning, distillation, and information-maximization, which all fall under the banner of discriminative SSL, have experienced a swift uptake in the area of computer vision. Consequently, within a span of three years, more than 50 unique general-purpose frameworks for discriminative SSL, with a focus on images, were proposed. In this survey, we review a plethora of research efforts conducted on image-oriented SSL, paying attention to best practices and useful software packages. While doing so, we discuss pretext tasks for image-based SSL, as well as techniques that are commonly used in discriminative SSL. Lastly, to aid researchers who aim at contributing to image-focused SSL, we outline a number of relevant research directions.

1 Introduction

The remarkable feature extraction capabilities of deep neural networks (DNNs) enabled their effective utilization in numerous visual tasks. Although the core building blocks that are in common use today were already proposed two decades ago LeCun et al. (1998a), DNNs only became the go-to models after the introduction of AlexNet Krizhevsky et al. (2012), a DNN architecture that was able to obtain exceptional results for the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015) that took place in 2012, by leveraging vast amounts of computational resources (at that time) and large amounts of labeled data. Since then, the availability of standardized datasets in the image domain such as MNIST (LeCun et al., 1998b), CIFAR (Krizhevsky & Hinton, 2009), SVHN (Netzer et al., 2011), COCO (Lin et al., 2014), and ImageNet enabled standardized experimentation, with these datasets acting as catalysts for major advancements in the area of supervised learning. Starting with AlexNet, the classification accuracy of DNNs on ImageNet improved year after year thanks to better and novel architectural designs (e.g., VGG (Simonyan & Zisserman, 2015), ResNet (He et al., 2016), InceptionNet (Szegedy et al., 2015; 2016), ViT (Dosovitskiy et al., 2020)), augmentation techniques, optimizers, and activation functions, as well as methods for smoother training (Loshchilov & Hutter, 2017; Yun et al., 2019; Ioffe & Szegedy, 2015; Kingma & Ba, 2014; Clevert et al., 2015).

Unfortunately, not all datasets come with an abundance of labeled training data. In order to overcome this hurdle and to facilitate the application of DNNs to smaller datasets, transfer learning was introduced and soon became the dominant method to convey knowledge across image datasets (Tan et al., 2018). Although transfer learning enables the usage of DNNs for smaller datasets thanks to features extracted from larger datasets, models trained in this way are known to be brittle and sensitive to small changes in the data (Jain et al., 2022) due to the use of supervised pre-training. Furthermore, shortcomings of supervised learning also became apparent when improvements obtained with these methods came to a halt in recent years (see Figure 1 for top-1 accuracy on ImageNet), thus calling for research efforts that go beyond the use of supervised



(a) ImageNet top-1 accuracy

(b) Interest over time for different learning paradigms

Figure 1: (left) ImageNet top-1 accuracy for DNNs proposed between 2012 - 2022 and (right) interest over time for three popular learning paradigms between 2004 - 2022, as measured with Google Trends.

learning (Zisserman, 2018). In order to overcome the limitations of supervised learning, countless studies investigated the line of unsupervised learning, which aims at enabling robust feature extraction through the training of models without label information (Celebi & Aydin, 2016). Unfortunately, results obtained by these methods on image datasets, until recently, fell short in recent years (Noroozi & Favaro, 2016; Pathak et al., 2016), while the use of self-supervised methods in the field of natural language processing (NLP) achieved state-of-the-art results, compared to supervised learning techniques Devlin et al. (2018); Radford et al. (2019).

As mentioned above, the field of NLP enjoyed the success of self-supervised models over supervised ones earlier than the field of computer vision, with models such as `BERT`, `GPT`, and their variants achieving state-of-the art results (Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020). One reason which explains the success of SSL in NLP is the abundance of unlabeled text data, such as books, online websites, and blogs (Chen et al., 2017; Hamilton et al., 2017), which prompted researchers to investigate SSL over supervised training. Another reason that explains their success, as discussed by He et al. (2020), is the fundamental difference between the signal space of NLP and the signal space of computer vision, given that language data are discrete and structured (i.e., words), whereas image data are high dimensional, continuous, and unstructured. Nevertheless, we can state that the success of SSL in the field of NLP prompted the computer vision community to put more investigative efforts into this learning paradigm.

In order to alleviate issues regarding label requirements, as well as to enable robust feature extraction, self-supervised learning in computer vision emerged as a method for extracting robust features from unlabeled data using the properties of images themselves (He et al., 2020; Chen et al., 2020b). The idea behind SSL is straightforward: devise an experimental setting in which the task that provides the supervisory signal can be solved without human annotation and then train DNNs to solve it.

Note that the description provided above for SSL also covers a number of additional approaches including autoencoders (Gogna & Majumdar, 2016), generative models, and clustering-based methods that leverage self-labeling (Caron et al., 2018), and that these approaches also fall into the category of unsupervised learning (since human annotation is not necessary). Furthermore, most of the training routines described in this manuscript also use the term "self-supervised learning" interchangeably with "representation learning" when supervision is provided by the data, while representation learning is described by Bengio et al. (2013) as "learning representations of the data that make it easier to extract useful information when building classifiers or other predictors", irrespective of the supervisory nature of the learning methodology. So, how did "self-supervision" became such a popular term in recent years?

Resurgence of the term "self-supervised learning" in computer vision – Beyond a number of niche use cases such as image colorization (Larsson et al., 2017), image inpainting (Yang et al., 2017), and puzzle-solvers (Trinh et al., 2019) that explicitly use self-supervision, the term "self-supervised learning" was previously not employed to describe many techniques. Furthermore, compared to other learning paradigms, the use of SSL was not popular until recently (see Figure 1). In fact, research efforts that are now considered to be pioneers in SSL and that are used for SSL benchmarking, such as `Deep Cluster` (Caron et al., 2018), `InsDist` (Wu et al., 2018b), `CPC` (Oord et al., 2018), and `Local Aggregation` (Zhuang et al., 2019), were published as unsupervised training methods, explicitly distancing themselves from SSL.

The resurgence of interest in self-supervision and the re-branding of corresponding methodologies can be attributed to the popularization of the term by both authoritative researchers and tech giants in the field

between 2018 and 2020 (Zisserman, 2018; Efros, 2019; Bachman, 2019; LeCun & Misra, 2020; Chen, 2020; Howard, 2020). The reason for this re-branding is straightforward: most of the tasks discussed above that fell under the banner of unsupervised learning were deemed misleading, since the training was not completely unsupervised. Instead, the supervision was provided by the data themselves, without explicit human labeling (Zisserman, 2018; LeCun, 2019). As a result of this re-branding, while most papers published before 2020 use unsupervised learning to describe their work, those that are published after 2020 use the description self-supervised learning, hence the conflict between the use of the two terms.

An interesting moment in this timeline, and the one that furthered the popularity of the term SSL, is the revision by Yann LeCun of his now-famous cake analogy from NeurIPS-16, during a talk he gave at ISSCC-19 and later at AAAI-20 (LeCun, 2020): "If intelligence is a cake, the bulk of the cake is ~~unsupervised~~ *self-supervised* learning, the icing on the cake is supervised learning, and the cherry on the cake is reinforcement learning" (LeCun, 2016).

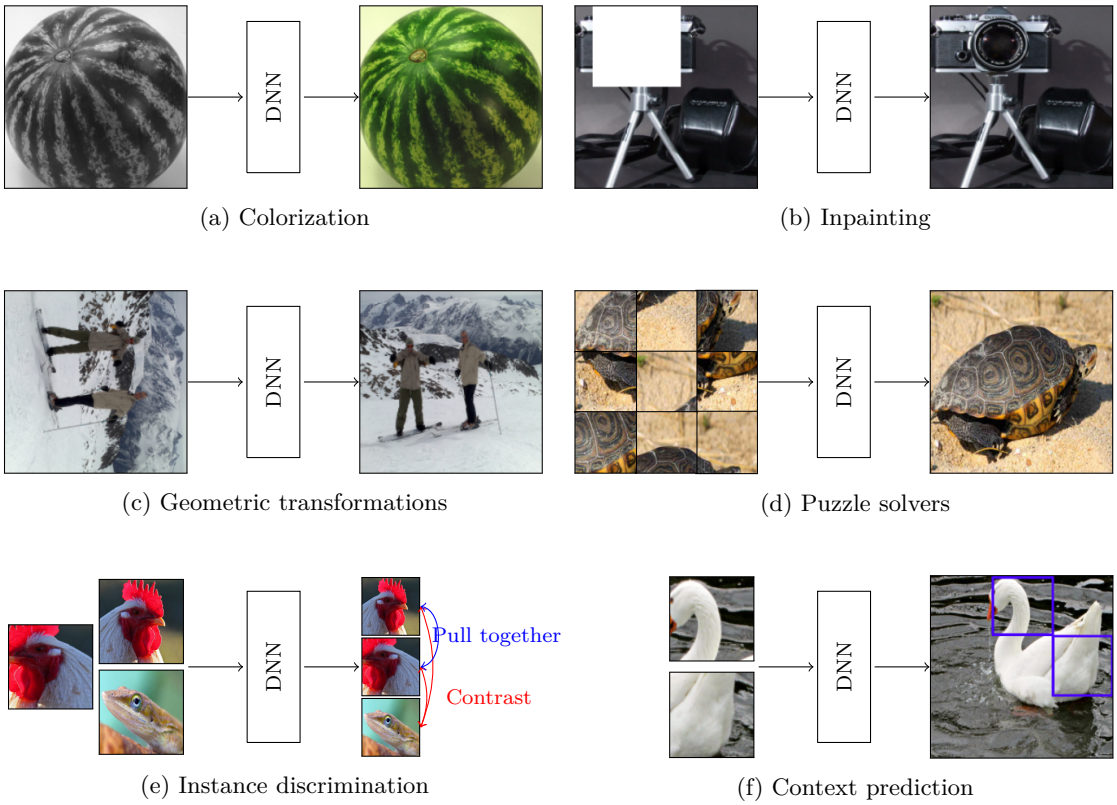
In summary, we can say that self-supervised learning refers to a recently popularized learning paradigm, encompassing predictive tasks where the supervisory signal is provided by the data, without relying on the explicit use of human labels.

Generative and discriminative SSL – In general, self-supervision can be grouped into two categories: generative and discriminative (Doersch et al., 2015). In generative self-supervision, the task is to build appropriate distributions over a collection of data while operating in the pixel space. Some common criticism of generative self-supervision is that it is computationally expensive, does not work well with high-resolution images, and that it may be superfluous for representation learning (Chen et al., 2020b; Grill et al., 2020). Typical models relying on this kind of self-supervision are autoencoders and generative models (Kingma & Welling, 2013; Vincent et al., 2008; Goodfellow et al., 2020).

In discriminative self-supervision, on the other hand, the task is to learn good representations of the data in order to perform a specified pretext task (which we will explain shortly) that does not require a human annotation effort (Doersch et al., 2015). Discriminative self-supervision is similar to supervised learning in the sense that the objective function is often a scoring function that evaluates the discriminative power of learned representations. Most of the SSL frameworks we will cover in this manuscript refer to the works of Becker & Hinton (1992) and Bromley et al. (1993) as the earliest research efforts that use discriminative self-supervision in the form it is used nowadays, with the above research efforts investigating representation alignments across different inputs.

Purpose of this survey – Thanks to the excellent results obtained by SSL in vision, numerous SSL frameworks were proposed within the span of a couple of years. Although most of these frameworks are often specialized in nature, addressing a select number of tasks (such as depth estimation, face recognition, remote sensing, and pose estimation), we could trace their origin to roughly 50 general-purpose SSL frameworks that are applicable to images. Even though several in-depth surveys are available on the topic of image-based contrastive SSL (Albelwi, 2022; Khan et al., 2022), due to the fast-paced nature of research in SSL, they do not cover recent non-contrastive SSL methods that transformed the field. As such, a major goal of this survey is to cover the aforementioned image-oriented frameworks for discriminative SSL which benefited from a tremendous research and development efforts in recent years, hereby presenting a concise and aggregate work to readers who take an interest in this field.

In Section 2, we describe popular pretext tasks for self-supervision, subsequently detailing a number of relevant technical concepts that are commonly used in Section 3. Diving deeper into SSL as it is used nowadays for image-related tasks, in Section 4, we cover recently proposed SSL frameworks for image-based training in a chronological order and discuss methods of evaluation in Section 5. In Section 6, we discuss methods of SSL evaluation and relevant libraries, repositories, and publicly available implementations that aim at assisting researchers. Finally, in Section 7, we explain shortcomings of SSL, identify open problems, and conclude our survey.



(a) Colorization (b) Inpainting

(c) Geometric transformations (d) Puzzle solvers

(e) Instance discrimination (f) Context prediction

Figure 2: Illustrations of various image-based pretext tasks for self-supervised learning.

2 Pretext tasks for self-supervised learning

The image domain allows a number of unique pretext tasks that enable self-supervision. Below we describe the most popular ones and illustrate them in Figure 2.

Image colorization – Automated colorization of grayscale images is a line of research that was investigated even before the widespread usage of DNNs (Luan et al., 2007; Charpiat et al., 2008). However, the availability of large-scale colored datasets such as ImageNet, combined with the versatility of DNNs, further strengthened the interest in high-quality image colorization, especially for the purpose of coloring historical pictures. Parallel to research efforts that aimed at increasing the quality of colorization, such as Cheng et al. (2015); Iizuka et al. (2016), the idea of using image colorization as a pretext task for representation learning was also investigated (Larsson et al., 2017; 2016; Zhang et al., 2016). Although this task alone was revealed to be too simple to force DNNs to learn complex representations (Caron et al., 2020), colorization is still used in tandem with other tasks to boost the effectiveness of SSL models.

Inpainting – The task of predicting a missing part of an image is referred to as image inpainting (Bertalmio et al., 2000). With the widespread usage of DNNs, inpainting problems also found numerous solutions (Yang et al., 2017; Yu et al., 2019). One such solution, and the one that allows for the use of SSL, is proposed by Pathak et al. (2016), leveraging context encoders that aim at inpainting large parts of images that are missing, forcing models to learn the image context.

Geometric transformations – Inspired by research efforts that bring together geometric transformations and neural networks (Kanazawa et al., 2016; Rocco et al., 2017), and taking advantage of image-based datasets that almost always contain upright images, Gidaris et al. (2018) proposed the idea of predicting image rotations as a method of self-supervision. Following the success of this method, other types of geometric transformations were proposed by Novotny et al. (2018); Zhang et al. (2019); Chen et al. (2019).

Puzzle solvers – A unique image-based task that can be formulated in a SSL setting is solving a jigsaw puzzle (Noroozi & Favaro, 2016), where the goal is to correctly predict the relative location of nine puzzle pieces. This unusual pretext task, as well as a number of derivations, is employed in support of a variety of tasks, including domain generalization (Carlucci et al., 2019), generation of image embeddings (Trinh et al., 2019), image retrieval (Pang et al., 2020), and auxiliary learning (Li et al., 2021b).

Instance discrimination – Given differently augmented views (i.e., instances) originating from one image, instance discrimination refers to the idea of recognizing these views as originating from the same image, while discriminating any other image with a different origin (Wu et al., 2018b). Different from the previously described pretext tasks which achieve representation learning as a by-product of the optimization objective, instance discrimination optimizes for representation learning by directly matching the representations of similar images while contrasting the representations of dissimilar ones. In this context, images that are contrasted to similar ones are called negative samples (e.g., the gecko image in Figure 2e). The main idea behind representation matching between similar images and contrasting different images is to help DNNs learn representations that are invariant to commonly used image transformations, since most of these transformations do not alter the visual semantics (Misra & Maaten, 2020). The origins of this approach can be traced back to the research efforts presented in Hadsell et al. (2006), Sohn & Lee (2012), and Hui (2013).

Others – Apart from the mainstream pretext tasks described above, there are a number of unique tasks that do not fit into one of the above categories. In what follows, we list a number of popular ones: the split-brain approach which tries to predict a subset of image channels from other channels (Zhang et al., 2017), a feature consistency method involving synthetic images (Ren & Lee, 2018), context prediction (Doersch et al., 2015), adversarial feature learning (Donahue et al., 2016; Donahue & Simonyan, 2019), exemplar networks (Dosovitskiy et al., 2014b), pixel-based generative training (Chen et al., 2020a), and object counting (Noroozi et al., 2017).

Effectiveness of pretext tasks – Given the abundance of pretext tasks for self-supervision, which of these tasks enable networks to learn the most useful representations? Although there is no clear answer to this question, ever since the works of Dosovitskiy et al. (2014a), Wu et al. (2018b), and Oord et al. (2018), instance discrimination was established as the dominant pretext task for image-based SSL (He et al., 2020; Grill et al., 2020; Chen et al., 2020b), thanks to superb results achieved using this type of self-supervision.

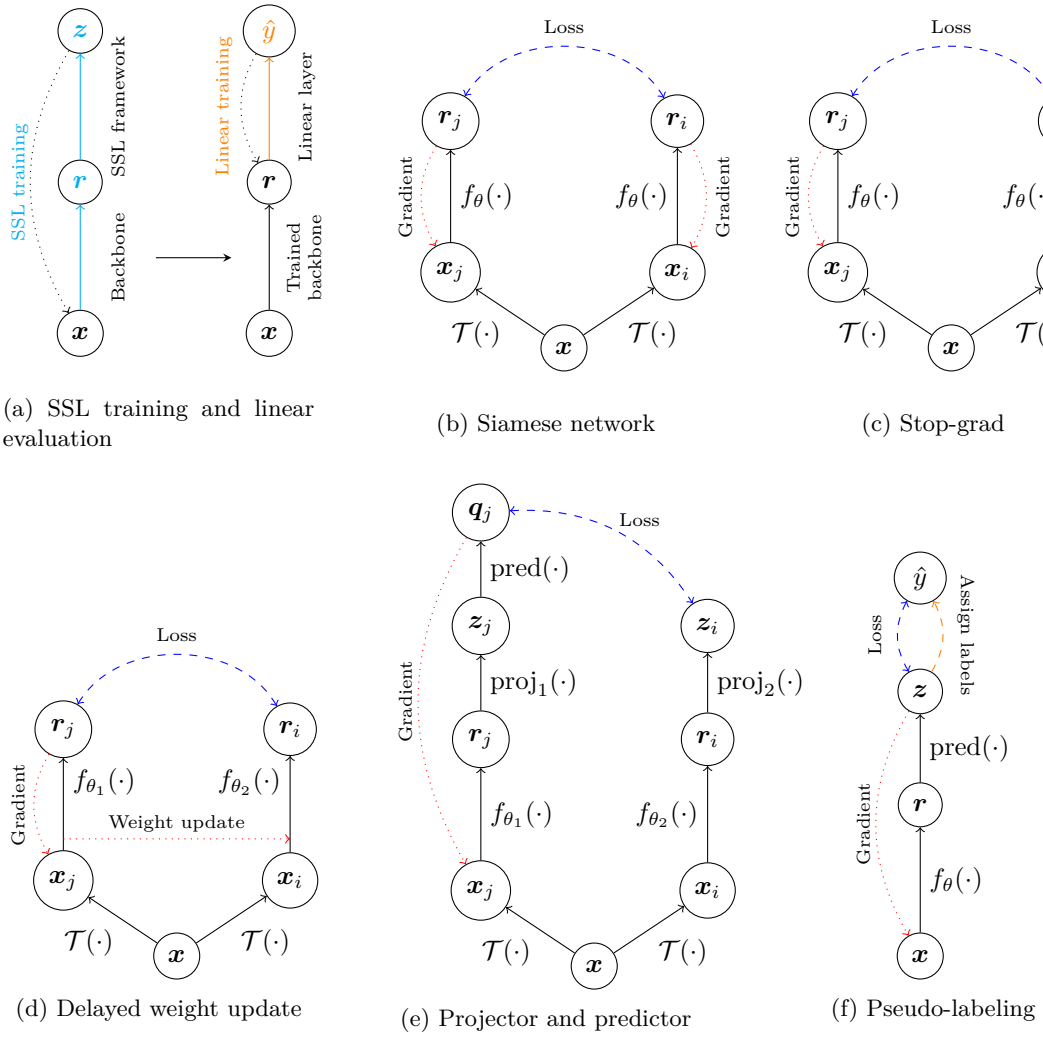
3 Important concepts in self-supervised learning

In this section, we briefly describe a number of commonly used concepts that are relevant to the forthcoming SSL frameworks. Although these concepts were key elements of early individual SSL frameworks, newer frameworks make use of a mixture of them.

Notation – For clarity, we briefly detail the notation used to describe several core SSL concepts. Given an image $\boldsymbol{x} \in \mathbb{R}^p$ and its categorical association $\boldsymbol{y} \in \mathbb{R}^M$ sampled from a dataset $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}$, with $y_c = 1$ and $y_m = 0\,, \forall\, m \in \{0, \ldots, M\} \backslash \{c\}$, let $f_\theta(\cdot)$ be an encoder (i.e., a feature extractor) that maps an image augmented with a stochastic augmentation function $\mathcal{T}(\cdot)$ to a set of features $\boldsymbol{r} \in \mathbb{R}^k$ using a neural network with parameters θ. These features can then be mapped onto a set of projections $\boldsymbol{z}$ and predictions $\boldsymbol{q}$ using the proj$(\cdot)$ and pred$(\cdot)$ functions, respectively. In this context, projectors and predictors are simply multi-layer perceptrons (MLP).

Backbone network – In the context of SSL, the term "backbone" refers to the feature extractor(s) (i.e., $f_\theta(\cdot)$) that are trained with SSL frameworks. Typically, a backbone network is a task-agnostic DNN (e.g., a ResNet-50 without the final fully connected layer). Majority of the frameworks we will cover use a variant of ResNet (e.g., vanilla ResNet-50, ResNext, or Wide ResNet) and, very recently, vision transformers as the backbone.

SSL training and evaluation – In traditional supervised learning, the feature extractor (e.g., convolutional layers) and the predictor (e.g., linear layers that map features to classes) are trained at the same time. However, SSL is only concerned with the training of the feature extractor. After the SSL training is complete, the linear layer that maps the features to classes is trained separately.



(a) SSL training and linear evaluation

(b) Siamese network

(c) Stop-grad

(d) Delayed weight update

(e) Projector and predictor

(f) Pseudo-labeling

Figure 3: Illustrations of some of the important concepts related to SSL described in Section 3.

In Figure 3a, we provide a simplified illustration of (left) SSL training and (right) linear evaluation. SSL frameworks are placed on top of backbone networks and are trained in conjunction with the backbone. After the SSL training is complete, the framework is discarded and only the trained backbone is used. Note that this backbone is merely a feature extractor. Then, depending on the problem at hand, a new layer that maps features to classes is initialized and trained. It is crucial to understand that the SSL training is only concerned with the quality of features obtained from the feature extractor. As such, the majority (if not entirety) of the forthcoming concepts as well as frameworks tackle feature extractor training. Nevertheless, for the sake of completeness, in Section 5, we will also describe evaluation methods.

Siamese networks – A form of dual-backbone networks called Siamese networks (Bromley et al., 1993) consisting of two identical neural networks that share the same set of weights (see Figure 3b) are popular architectures for SSL. Although this type of networks was useful in solving a variety of problems (Chopra et al., 2005b; Bertinetto et al., 2016; Chicco, 2021), in the context of SSL, they are mostly employed to achieve consistency between representations when, for example, two instances of the same image are provided.

Apart from Siamese networks, a majority of SSL frameworks use dual backbones that may not share weights due to recently discovered beneficial properties. In such cases, the weights of one model are updated via backpropagation, while the weights of the other model can be updated using a variety of techniques which we discuss next.

Stop-grad – Siamese networks generally propagate errors from both branches after the loss calculation. As illustrated in Figure 3c, the term "stop-grad" refers to stopping the gradient flow from one branch of a dual-backbone network, while allowing this gradient flow to alter the weights of the other branch (Chen & He, 2021).

Delayed weight updates – Assume a Siamese-like dual-backbone network where one branch is called the teacher and the other one the student. However, different than the Siamese architecture, weights of these models are not shared. In this scenario, delayed-weight updates refer to the idea of propagating the error through only one branch via backpropagation and updating the trainable parameters of the other branch via predetermined rule (see Figure 3d). Popular implementations of this operation are *Mean Teacher* (Tarvainen & Valpola, 2017), *momentum encoding* (He et al., 2020), and *exponential moving average* (Grill et al., 2020).

Projection and prediction MLPs – The usage of multi-layer perceptrons in the form of projection and prediction heads following a feature extractor (e.g., a dual backbone) is acknowledged as a powerful technique that greatly improves the effectiveness of SSL methods (Chen et al., 2020d). We visualize this technique in Figure 3e, as implemented in `BYOL` framework (Grill et al., 2020). Note that this visualization illustrates an asymmetric architecture but the asymmetry is not a necessity for projection/prediction MLPs.

Noise-contrastive estimation – A contrastive loss is a loss that has a low value when the two input images are similar and a large value when they are dissimilar (Chopra et al., 2005a; Hadsell et al., 2006). A fundamental loss that enabled contrastive training for image-based SSL has been InfoNCE(Sohn, 2016; Oord et al., 2018), which is a modification of Noise-Contrastive Estimation (NCE) (Gutmann & Hyvärinen, 2010). Following Chen et al. (2020b), InfoNCE can be defined using $2n$ instances of n images in a single batch: $\mathtt{x} = [T(\boldsymbol{x}_1), T(\boldsymbol{x}_1), \ldots, T(\boldsymbol{x}_n), T(\boldsymbol{x}_n)]$, with $T(\cdot)$ denoting a stochastic image augmentation function. In this scenario, the InfoNCE loss for a single positive pair is defined as follows:

$$\mathcal{L}_{\text{InfoNCE}}(\mathtt{x}_{\{i,j\}}) = -\log \frac{\exp(\text{sim}(\boldsymbol{r}_i, \boldsymbol{r}_j))}{\sum_{k=0}^{2n} \mathbb{1}_{\{k \neq i\}} \exp(\text{sim}(\boldsymbol{r}_i, \boldsymbol{r}_k))}\,, \tag{1}$$

where $f(\mathtt{x}_i) = \boldsymbol{r}_i$ denotes the feature representation of the ith data point and where $\text{sim}(\cdot, \cdot)$ represents cosine similarity.

Negative samples – The InfoNCE loss defined previously aims at minimizing the cosine similarity between representations of two augmentations of the same image, while maximizing the same metric across different images. In such cases, the (different) images are referred to as *negative samples* (Chen et al., 2020b). This concept, which has been the focus of many research efforts (which we will discuss later on), will be particularly relevant for contrastive SSL (He et al., 2020).

Memory bank – Given a set of n images, $\mathtt{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, a memory bank refers to the simple idea of storing the corresponding image representations, as computed with $f_\theta(\mathtt{x}) = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n]$, and to subsequently use this memory bank for various tasks (for example, to use the obtained image representations as negative samples in InfoNCE) (Wu et al., 2018a; He et al., 2020).

Pseudo-labeling – A number of SSL methods discussed below employ pseudo-labeling strategies to enable self-supervision (Caron et al., 2018; Asano et al., 2019). Such approaches can be visualized as shown in Figure 3f, where a label is assigned to an image based on its feature representation (through the use of, for example, K-means clustering) and where that label is then used to calculate a loss.

4 Self-supervised learning frameworks

Although contrastive self-supervision was the most prominent approach for self-supervision in the past, other approaches relying on clustering, distillation, and, very recently, information-maximization were shown to achieve comparable and, in some cases, vastly better results. In what follows, we give an overview of SSL frameworks in the four aforementioned categories. Note that a number of SSL frameworks leverage techniques that belong to multiple categories and may thus fall into more than one category. For these SSL frameworks, we adopt the designation used by their creators. Since most of the frameworks are known by their acronyms, we use their abbreviated names in the main text and provide their full names in the appendix.

Table 1: SSL frameworks that rely on **clustering**-based self-supervision and their unique properties.

SSL framework	Proposed by	Unique property
`Deep Cluster`	Caron et al. (2018)	Avoids trivial solutions for clustering-based SSL
`Local Aggregation`	Zhuang et al. (2019)	Local aggregation metric for soft cluster assignments
`Deeper Cluster`	Caron et al. (2019)	Integrates rotation-based SSL into clustering
`SeLa`	Asano et al. (2019)	Improves `Deep Cluster` with the Sinkhorn-Knopp algorithm
`SCAN`	Van Gansbeke et al. (2020)	Decouples feature learning and clustering using a two-step approach
`Deep Cluster-v2`	Caron et al. (2020)	Incorporates various SSL improvements into `Deep Cluster`
`SeLa-v2`	Caron et al. (2020)	Incorporates various SSL improvements into `SeLa`
`Swav`	Caron et al. (2020)	Online clustering with consistency across assignments
`ODC`	Zhan et al. (2020)	Converts `Deep Cluster` into an online method
`CoKe`	Qian et al. (2022)	Improves the clustering phase with an online constrained k-means method
`Self-Classifier`	Amrani & Bronstein (2021)	Single-stage end-to-end clustering combined with contrastive learning

4.1 Clustering

Self-labeling via clustering is one of the most straightforward ways to achieve self-supervision, with clustering being one of the most popular methods for unsupervised learning (Bishop, 2006). For neural networks, the usage of clustering-based methods for training can be traced back to the seminal works of Coates et al. (2011), Coates & Ng (2012), and Yang et al. (2016), which paved the way for the employment of such methods for SSL. Unfortunately, clustering-based methods have to solve a number of well-documented issues such as: (1) offline training that prevents their usage for large-scale data, (2) large clusters dominating the majority of the labels or small clusters leading to extremely granular labels, (3) empty clusters, (4) needing the knowledge on the number of clusters beforehand, and (5) trivial solutions where all data are gathered in a single cluster which causes the network to collapse (Xu et al., 2004; Joulin et al., 2016). Since these issues are fundamental problems of clustering, all of the clustering-based SSL methods have to tackle these problems in their own unique way when trying to perform self-supervision.

The pioneering work of Caron et al. (2018) put forward `Deep Cluster`, one of the first clustering-based SSL methods that achieves results comparable to supervised models. This method solves the issues listed above with an offline training approach and by forcing a uniform distribution across clusters, both of which limits the usage of `Deep Cluster`. Following that, getting rid of the tricks applied in `Deep Cluster` became the primary focus of a number of subsequent studies, leading to improved clustering-based SSL methods such as `SeLA` (Asano et al., 2019), `Online Deep Cluster` (Zhan et al., 2020), and `Self-Classifier` (Amrani & Bronstein, 2021). `SeLa` tackles the issue of model collapse by incorporating a more principled loss using the Sinkhorn-Knopp algorithm (Cuturi, 2013). `Online Deep Clustering` on the other hand addresses the aforementioned offline training issue to enable online training for large datasets.

Conversely, Van Gansbeke et al. (2020) argue that an end-to-end approach with online training may lead to various problems and propose an approach called `SCAN` that replaces the use of K-Means for the purpose of clustering with the use of an advanced neighbor search. When it comes to the state-of-the-art, the clustering-based method proposed in Caron et al. (2020), known as `Swav`, which also leverages a number of contrastive elements, is currently considered to be the most stable and accurate approach. Table 1 provides a summarizing overview of several clustering-based SSL methods, detailing their unique traits.

4.2 Contrastive learning

Contrastive learning with the InfoNCE (or an extension of it) is the most popular approach for self-supervision and also the one that received the most research contributions in the past years. Contrastive methods can be traced back to the works of Bromley et al. (1993) and Chopra et al. (2005b), but in terms of modern usage of SSL, Wu et al. (2018b) and Oord et al. (2018) popularized this line of research by proposing `InstDist` and `CPC`, respectively. Hjelm et al. (2018) and Bachman et al. (2019) investigated different ways to measure representation quality for contrastive learning and proposed `DIM` and `AMDIM` respectively, while Tian et al. (2020a) extended contrastive learning for multiple sensory inputs with `CMC`. After the aforementioned works contrastive SSL attracted a significant research interest but it was the groundbreaking results obtained with `MoCo` which used memory banks with delayed weight updates that put contrastive SSL really into the spotlight (He et al., 2020). Shortly after, Chen et al. (2020b) proposed `SimCLR` and with it, further improved the state-of-the-art wit the help of projection heads and strong augmentations and cemented

Table 2: SSL frameworks that rely on **contrastive learning**-based self-supervision and their unique properties.

SSL framework	Proposed by	Unique property
InstDist (NPID)	Wu et al. (2018b)	Non-parametric softmax calculation
CPC	Oord et al. (2018)	Usage of InfoNCE loss across multiple tasks
DIM	Hjelm et al. (2018)	Measures representation quality with two novel losses (MINE and NDM)
CPC-v2	Henaff (2020)	Improves CPC architecture and training
AMDIM	Bachman et al. (2019)	Extends DIM for mixture-based representations
CMC	Tian et al. (2020a)	Information-maximization across different sensory views
MoCo	He et al. (2020)	SSL with momentum encoder and memory bank
PIRL	Misra & Maaten (2020)	Contrastive learning with jigsaw puzzles
SimCLR	Chen et al. (2020b)	Usage of projection heads and new augmentations
MoCo-v2	Chen et al. (2020d)	Improves MoCo with the design of SimCLR
InfoMin	Tian et al. (2020b)	InfoMin principle and evaluation of augmentations
SimCLR-v2	Chen et al. (2020c)	Improves SimCLR with memory bank and deeper projector MLPs
PCL & PCL-v2	Li et al. (2020)	Formulates contrastive learning with clustering using EM
PIC	Cao et al. (2020)	One-branch parametric instance classification
DCL	Chuang et al. (2020)	Negative sample section with a debiased contrastive objective
InterCLR	Xie et al. (2022a)	Inter-image invariance for contrastive learning
G-SimCLR	Chakraborty et al. (2020)	SimCLR with negative sample selection using pseudo-labels
MocHi	Kalantidis et al. (2020)	Hard negative image mixing approach
ReLIC	Mitrovic et al. (2020)	Imposes invariance constraints during SSL training
AdCo	Hu et al. (2021)	Mixes self-trained negative adversaries into SSL
DenseCL	Wang et al. (2021b)	Dense contrastive loss for SSL
PixPro	Xie et al. (2021c)	PixContrast and PixPro losses for contrastive SSL
MoCo-v3	Chen et al. (2021)	Improves MoCo-v2 with symmetrized loss and without a memory bank
CLSA	Wang & Qi (2022)	Usage of stronger augmentations for contrastive learning
NNCLR	Dwibedi et al. (2021)	Nearest-neighbors as positive samples in contrastive loss
MoBY	Xie et al. (2021b)	Combines design principles of MoCo and BYOL for transformers
DNC	Tian et al. (2021a)	Alternation of contrastive learning and clustering-based hard negative mining
ReSSL	Zheng et al. (2021)	Maintains the relational consistency between different instances of images
UniGrad	Tao et al. (2022)	Unifies contrastive-learning, distillation, and information-maximization
ReLIC-v2	Tomasev et al. (2022)	Improves ReLIC with inductive biases to learn more informative representations
SimCo	Zhang et al. (2022a)	Simplifies MoCo with momentum removal
SimMoCo	Zhang et al. (2022a)	Simplifies MoCo with dictionary removal
UniVIP	Li et al. (2022b)	Scene-based SSL based on similarity, correlation, and discrimination
SMoG	Wang et al. (2023)	Replaces instance constrastive learing with group contrastive learning
MosRep	Wang et al. (2023)	Proposes a new data augmentation strategy that enriches the backgrounds of crops

the importance of contrastive self-supervision as a learning paradigm. Incorporating the enhancements of SimCLR into MoCo, Chen et al. (2020d) proposed MoCo-v2 and showed that there still exists a large margin for improvement. Chen et al. (2021) later introduced a third version of MoCo, exploring the usage of vision transformers as backbones. The reliable design of MoCo and its improved versions were the foundation of many subsequent contrastive SSL frameworks, such as AdCo (Hu et al., 2021), MocHi (Kalantidis et al., 2020), and DenseCL (Wang et al., 2021b).

While the above architectures mostly use dual backbones, Cao et al. (2020) proposed PIC and demonstrated the viability of architectures with a single-branch backbone for contrastive learning. Kalantidis et al. (2020) experimented with hard negative samples for improving the effectiveness of contrastive learning and Wang & Qi (2022) demonstrated the usefulness of stronger augmentations. After the success of Moco-v2 and Moco-v3, and with the increased availability of unique SSL methods, frameworks like G-SimCLR (Chakraborty et al., 2020), MoBY (Xie et al., 2021b), SimCo, and SimMoCo (Zhang et al., 2022a), which combine multiple SSL methods into a single one, gained traction. More recently, SSL frameworks such as UniGrad (Tao et al., 2022) and UniVIP (Li et al., 2022b) claim to combine four self-supervision methodologies (clustering, contrastive, distillation, and information-maximization) into a single framework and to unify discriminative SSL training.

Although contrastive methods garnered more attention than clustering-based methods, they are also subject to a similar problem that needs to be mitigated: network collapse (Jing et al., 2021). Contrastive methods prevent complete collapse of a network through the use of negative samples. However, Hua et al. (2021) surprisingly demonstrated that contrastive SSL frameworks can suffer from another type of collapse, namely dimensional collapse, wherein representations collapse into a low-dimensional manifold. Given the importance of negative samples in preventing collapse in contrastive SSL, understanding the effects of negative samples and finding better sampling techniques became an active research topic shortly after (Chuang et al., 2020; Robinson et al., 2020; Zhang et al., 2022a). A summarizing overview of several contrastive SSL frame-

Table 3: SSL frameworks that rely on **distillation**-based self-supervision and their unique properties.

SSL framework	Proposed by	Unique property
BYOL	Grill et al. (2020)	Avoids trivial solutions through network asymmetry
SimSiam	Chen & He (2021)	SSL with simple Siamese networks without negative samples
OBoW	Gidaris et al. (2021)	Online bag-of-visual-words for SSL
DirectPred	Tian et al. (2021b)	Adjusts linear predictor with a gradient-free approach
SEED	Fang et al. (2021)	Knowledge distillation from large to small models
DisCo	Gao et al. (2021)	Combines contrastive and distillation learning for lightweight models
DINO	Caron et al. (2021)	Knowledge distillation with vision transformers
MSF	Koohpayegani et al. (2021)	Enhances BYOL by shifting the embeddings to be close to the mean of its instances
ORL	Xie et al. (2021a)	Utilizes BYOL for object-level training
CEB	Lee et al. (2021)	Measures the amount of compression in the learned representations
BINGO	Xu et al. (2021)	Distillation-based SSL for small-scale models
SEM	Lavoie et al. (2023)	Employs simplistic embeddings to map unnormalized representations onto simplices

Table 4: SSL frameworks that rely on **information-maximization**-based self-supervision and their unique properties.

SSL framework	Proposed by	Unique property
WMSE	Ermolov et al. (2021)	Whitening Mean Squared Error loss for information-maximization
Barlow Twins	Zbontar et al. (2021)	SSL with redundancy reduction
VicReg	Bardes et al. (2021)	Variance-invariance-covariance regularization for avoiding collapse
TWIST	Wang et al. (2021a)	Theoretically explainable TWIST loss that avoids collapse
TLDR	Kalantidis et al. (2021)	Improves Barlow Twins with TLDR encoder
ARB	Zhang et al. (2022b)	Aligns feature representations with nearest orthonormal basis

works can be found in Table 2 and benchmarks for those frameworks on ImageNet can be found in Table 10 and Table 11.

4.3 Distillation

Can the collapse of networks be prevented without the use of self-labeling or of a contrastive loss that relies on negative samples? Through an asymmetric framework called BYOL, Grill et al. (2020) demonstrated that neither of those techniques are necessary to achieve self-supervision when the proposed method relies on distillation (Hinton et al., 2015). The general idea behind distillation is to train a network (student) to predict representations of another one (teacher) (Tarvainen & Valpola, 2017). Shortly after the proposal of BYOL, Chen & He (2021) proposed SimSiam, a symmetric (Siamese) framework that neither uses negative samples nor uses clustering but leverages stop-grad and projection/prediction MLPs. This was followed by OBOW (Gidaris et al., 2021), in which the task is to reconstruct a bag-of-visual-words representation.

Similar to the trends witnessed for clustering and contrastive-learning, distillation-based SSL frameworks were experimentally combined with other frameworks in an attempt to obtain boosts in overall effectiveness. Frameworks such as DisCo (Gao et al., 2021) and MoBY (Xie et al., 2021b) merged multiple frameworks together, while others tried to improve the effectiveness of established methods, such as MSF (Koohpayegani et al., 2021) and ORL (Xie et al., 2021a), improving upon BYOL.

How do distillation methods avoid network collapse? Tian et al. (2020c) and Fetterman & Albrecht (2020) argued that methods that incorporate batch statistics into training (e.g., batch normalization) aid BYOL (and potentially other distillation-based methods) in preventing collapse, but this hypothesis was promptly refuted by Richemond et al. (2020). Recently, Li et al. (2022a) scrutinized SimSiam and found it to be highly sensitive to model size. Nevertheless, a definite answer to the way distillation-based SSL methods avoid collapse is not yet found. Table 3 provides a summarizing overview of several SSL frameworks that rely on distillation. ImageNet benchmarks for those frameworks can be found in Table 12.

4.4 Information-maximization

The fourth and final self-supervision category we cover is information-maximization, having as primary idea the maximization of the information conveyed by decorrelated embeddings. Such approaches come with a number of advantages, in particular, they neither require negative samples nor require an asymmetric architecture to avoid collapse. Instead, they completely rely on innovative loss functions to avoid collapse.

Table 5: Github repositories related to SSL, their maintainer, and purpose.

Repository name	Maintainer	Purpose
Awesome SSL	Independent	A comprehensive reading list for SSL
solo-learn	Independent (da Costa et al., 2022)	SSL frameworks, benchmarking, and model zoo
VISSL	Facebook (Goyal et al., 2021)	SSL frameworks, benchmarking, and model zoo
MMSelfSup	OpenMMLab (Contributors, 2021)	SSL frameworks, benchmarking, and model zoo
Lightly	Lightly.ai (Susmelj et al., 2020)	SSL frameworks and benchmarking
EasyCV	Alibaba (Contributors, 2022)	SSL frameworks and benchmarking
Unified SSL Benchmark	Microsoft (Wang et al., 2022a)	SSL frameworks and benchmarking

As a result, most of the frameworks that fall under this category can be characterized by by the novel loss function that is used.

Information-maximization as a method for self-supervision was put forward by Ermolov et al. (2021) and Zbontar et al. (2021), where the former proposed `W-MSE` loss, which constrains the batch samples to dissipate in a spherical distribution, and where the latter (`Barlow Twins`) aims at making the normalized cross-correlation matrix of the embedding vectors to be close to the identity matrix. Bardes et al. (2021) further improved the loss of `Barlow Twins` with the `VicReg` framework, proposing a loss based on variance, invariance, and covariance. Successor frameworks such as `TWIST` (Wang et al., 2021a), `TLDR` (Kalantidis et al., 2021), and `ARB` (Zhang et al., 2022b) followed the path paved by the previous frameworks and aim at improving the losses in different ways. Due to the complex nature of the losses used in information-maximization as a method for self-supervision, we refer the interested reader to the respective research papers underlying those frameworks. Table 4 provides a summarizing overview of several SSL frameworks that rely on information-maximization. ImageNet benchmarks for those frameworks can be found in Table 13.

5 Evaluating SSL models

As we briefly noted in Section 3, SSL frameworks we covered thus far are concerned with the training of feature extractors that can extract robust and useful features from images. Regardless, those feature extractors must be evaluated for a fair comparison of performance, which is what we will briefly cover in this section. Thus far in the SSL literature, trained models have been evaluated using two approaches: (1) linear evaluation (also called linear probing) and (2) K-nearest neighbor (KNN)-based evaluation through extracted features.

In linear evaluation, all trainable layers in the backbone are frozen after the SSL training and a newly introduced linear layer that maps features to predictions is trained. Another way to measure the performance of models is via KNN-based evaluation where a feature bank from the training data is generated. Then, predictions are assigned for test images based on KNN labels of the training data.

It is worth noting that the hyperparameters for the linear probing are tuned individually for most SSL frameworks, while in KNN-based evaluation, $k = 200$ is often used (following (Wu et al., 2018b)). Generally, the accuracy obtained with linear probing is higher than the accuracy obtained with KNN-based evaluation which is why it is often the selected method of evaluation when it comes to benchmarking. However, linear probing is more costly in terms of training time as well as more sensitive to hyperparameter selection.

6 Availability and comparability of SSL frameworks

Most of the frameworks covered in Section 4 perform experiments on both ImageNet and Pascal VOC (Everingham et al., 2007), thus enabling straightforward benchmarking and comparability. Not only that, many SSL frameworks come with implementations and trained models that are publicly available, contributing to speeding up research on SSL. For example, the availability and the straightforward adoptability of the `MoCo` framework enabled a number of follow-up studies that used the code of `MoCo` (Kalantidis et al., 2020; Hu et al., 2021; Wang et al., 2021b). For the SSL frameworks covered in this survey, in Table 8, we provide the availability of official implementation as well as trained models.

Apart from the availability of official implementations, the availability of third-party repositories also accelerated the adoption of SSL, enabling unified experimentation. Alas, not all third-party repositories are up

to date, and some of them have already been abandoned. In Table 5, we provide a number of useful SSL repositories that have been updated within the third quarter of 2022.

7 Conclusions and directions for future research

In this survey, in order to bring interested researchers up to speed with the field of SSL, we reviewed a number of general-purpose frameworks that use images for discriminative SSL training. In what follows, we highlight a number of directions for future research that are ripe for contribution.

Negative samples – One of the focal points for contrastive learning has been the usefulness of negative samples and their degree of difficulty. While some studies claim that hard negatives are useful for SSL compared to simple ones (Kalantidis et al., 2020), Xie et al. (2022a) argued that semi-hard negatives are more reliable and unbiased than hard negative samples. Nevertheless, the degree of usefulness of negative samples for contrastive learning is still a relevant research topic.

Fundamental questions for SSL – How do SSL frameworks avoid collapse without explicit guidance? What are the theoretical limitations of SSL? How do different types of loss affect the loss landscape? A number of fundamental questions regarding the limitations as well as the inner workings of SSL remain unanswered. While some research efforts tried to tackled some of the aforementioned questions (Garrido et al., 2022b), definite answers to most of these fundamental questions are still unknown (Chen et al., 2022).

Impact of augmentations – Although the impact of augmentations on SSL training is a topic that has thus far received more research attention than the aforementioned topics, conflicting results still make it an important future research direction. Furthermore, studies like Zhang et al. (2022c) and Wang et al. (2022b) demonstrate that we only scratched the surface of what is possible with SSL in terms of unique image augmentations.

Domain-specific SSL – The SSL frameworks covered in this survey mostly evaluate the effectiveness of models on the ImageNet. However, the results obtained for these datasets, which contain natural images, do not always generalize well to other datasets that contain different image modalities (e.g., medical imaging (Ramesh et al., 2022)). As a result, investigation of SSL frameworks that take advantage of the specifics of data in other domains is a highly relevant topic of research.

Calibration, interpretability, and adversarial robustness – Preliminary results indicate that models trained with SSL have different properties for robustness and interpretability as compared to models trained in a supervised fashion (Hendrycks et al., 2019; Zhong et al., 2022). Still, many of both the beneficial and detrimental properties of self-supervised training remain unknown.

Efficient SSL – SSL training requires a tremendous amount of computational power compared to supervised training. To give an example, (Chen et al., 2021) details that the training of `MoCo-v3` with a vision transformer backbone takes approximately 625 TPU days to train. As such, unfortunately, SSL has further exacerbated the computational requirements of DNN training which explains why almost all of the publications for frameworks discussed in this survey have at least an author with an industry affiliation (see Table 8 for a detailed breakdown on author affiliation) and not only that, the majority of the contributions ($> 80\%$) come from industry labs such as Facebook AI Research (14 frameworks), Microsoft Research (6 frameworks), DeepMind (6 frameworks), Google Research (4 frameworks), SenseTime Research (3 frameworks), and Huawei (2 frameworks). In order to mitigate this shortcoming of training cost, investigation of both efficient training and evaluation methods recently gained traction (Li et al., 2021a; Garrido et al., 2022a).

Rise of generative SSL – Generative SSL was dismissed in favor of discriminative training methods during the past couple of years, with contrastive methods holding the prime spot for research on SSL. In an unexpected turn of events, generative SSL frameworks for computer vision just recently demonstrated state-of-the-art results (Zhou et al., 2021; He et al., 2022; Xie et al., 2022b; Tian et al., 2023). Since this direction of research is new, we expect comparative and cooperative approaches using both generative and discriminative SSL to be highly valuable in the near future.

References

Saleh Albelwi. Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4):551, 2022.

Elad Amrani and Alex Bronstein. Self-supervised classification network. *arXiv preprint arXiv:2103.10994*, 2021.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

Philip Bachman. Going meta: learning algorithms and the self-supervised machine. In *Microsoft AI Podcast*, 2019. URL `https://www.youtube.com/watch?v=CSjWb3gcZJ4`.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 417–424, 2000.

Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*, pp. 850–865. Springer, 2016.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in Neural Information Processing Systems*, 6, 1993.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. *Advances in Neural Information Processing Systems*, 33:15614–15624, 2020.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2959–2968, 2019.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

M Emre Celebi and Kemal Aydin. *Unsupervised learning algorithms*. Springer, 2016.

Souradip Chakraborty, Aritra Roy Gosthipaty, and Sayak Paul. G-simclr: Self-supervised contrastive learning with guided projection via pseudo labelling. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pp. 912–916. IEEE, 2020.

Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. In *European Conference on Computer Vision*, pp. 126–139. Springer, 2008.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pp. 1691–1703. PMLR, 2020a.

Ting Chen. Advancing self-supervised and semi-supervised learning with simclr. In *Google AI Blog*, 2020. URL `https://ai.googleblog.com/2020/04/advancing-self-supervised-and-semi.html`.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12154–12163, 2019.

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, 2020b.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33:22243–22255, 2020c.

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*, 2020d.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Yubei Chen, Adrien Bardes, Zengyi Li, and Yann LeCun. Intra-instance vicreg: Bag of self-supervised image patch embedding. *arXiv preprint arXiv:2206.08954*, 2022.

Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proceedings of the IEEE international conference on computer vision*, pp. 415–423, 2015.

Davide Chicco. Siamese neural networks: An overview. *Artificial Neural Networks*, pp. 73–94, 2021.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005a.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005b.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in Neural Information Processing Systems*, 33:8765–8775, 2020.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast And accurate deep network learning by exponential linear units (Elus). *arXiv preprint arXiv:1511.07289*, 2015.

Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade*, pp. 561–580. Springer, 2012.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

EasyCV Contributors. Easycv. `https://github.com/alibaba/EasyCV`, 2022.

MMSelfSup Contributors. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. `https://github.com/open-mmlab/mmselfsup`, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23 (56):1–6, 2022. URL `http://jmlr.org/papers/v23/21-1155.html`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in Neural Information Processing Systems*, 27, 2014a.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in Neural Information Processing Systems*, 27, 2014b.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Alyosha Efros. The gelato bet. 2019. URL `https://people.eecs.berkeley.edu/~efros/gelato_bet.html`.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pp. 3015–3024. PMLR, 2021.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021.

Abe Fetterman and Josh Albrecht. Understanding self-supervised and contrastive learning with "bootstrap your own latent" (byol). In *Generally Intelligent AI Blog*, 2020. URL `https://generallyintelligent.ai/blog/2020-08-24-understanding-self-supervised-contrastive-learning/`.

Yuting Gao, Jia-Xin Zhuang, Ke Li, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Xing Sun. Disco: Remedy self-supervised learning on lightweight models with distilled contrastive learning. *arXiv preprint arXiv:2104.09124*, 2021.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. *arXiv preprint arXiv:2210.02885*, 2022a.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*, 2022b.

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

Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Pérez. Online bag-of-visual-words generation for unsupervised representation learning. *arXiv preprint arXiv:2012.11552*, 2021.

Anupriya Gogna and Angshul Majumdar. Semi supervised autoencoder. In *International Conference on Neural Information Processing*, pp. 82–89. Springer, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

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 Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

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.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

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

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Jeremy Howard. Self-supervised learning and computer vision. In *fast.ai Blog*, 2020. URL `https://www.fast.ai/2020/01/13/self_supervised/`.

Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1074–1083, 2021.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9598–9608, 2021.

Ka Yu Hui. Direct modeling of complex invariances for visual object features. In *International Conference on Machine Learning*, pp. 352–360. PMLR, 2013.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

Saachi Jain, Hadi Salman, Alaa Khaddaj, Eric Wong, Sung Min Park, and Aleksander Madry. A data-based perspective on transfer learning. *arXiv preprint arXiv:2207.05739*, 2022.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pp. 67–84. Springer, 2016.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.

Yannis Kalantidis, Carlos Lassance, Jon Almazan, and Diane Larlus. Tldr: Twin learning for dimensionality reduction. *arXiv preprint arXiv:2110.09455*, 2021.

Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3253–3261, 2016.

Adnan Khan, Sarah AlBarri, and Muhammad Arslan Manzoor. Contrastive self-supervised learning: a survey on different architectures. In *2022 2nd International Conference on Artificial Intelligence (ICAI)*, pp. 1–6. IEEE, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10326–10335, 2021.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pp. 577–593. Springer, 2016.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6874–6883, 2017.

Samuel Lavoie, Christos Tsirigotis, Max Schwarzer, Ankit Vani, Michael Noukhovitch, Kenji Kawaguchi, and Aaron Courville. Simplicial embeddings in self-supervised learning and downstream classification. In *International Conference on Learning Representations*, 2023.

Yann LeCun. Predictive learning. In *NIPS 2016*, 2016. URL `https://www.youtube.com/watch?v=0unt2Y4qxQo&t=1150s`.

Yann LeCun. I now call it "self-supervised learning". 2019. URL `https://www.facebook.com/yann.lecun/posts/10155934004262143`.

Yann LeCun. Self-supervised learning. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI, Invited Talk*, 2020. URL `https://drive.google.com/file/d/1r-mDL4IX_hzZLDBKp8_e8VZqD7fOzBkF/view`.

Yann LeCun and Ishan Misra. Self-supervised learning: The dark matter of intelligence. In *Facebook AI Blog*, 2020. URL `https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/`.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998a.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998b.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34:19538–19552, 2021.

Alexander C Li, Alexei A Efros, and Deepak Pathak. Understanding collapse in non-contrastive siamese representation learning. In *European Conference on Computer Vision*, pp. 490–505. Springer, 2022a.

Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021a.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Ru Li, Shuaicheng Liu, Guangfu Wang, Guanghui Liu, and Bing Zeng. Jigsawgan: Auxiliary learning for solving jigsaw puzzles with generative adversarial networks. *IEEE Transactions on Image Processing*, 31: 513–524, 2021b.

Zhaowen Li, Yousong Zhu, Fan Yang, Wei Li, Chaoyang Zhao, Yingying Chen, Zhiyang Chen, Jiahao Xie, Liwei Wu, Rui Zhao, et al. Univip: A unified framework for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14627–14636, 2022b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft Coco: Common Objects In Context. In *Proceedings of the IEEE European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Qing Luan, Fang Wen, Daniel Cohen-Or, Lin Liang, Ying-Qing Xu, and Heung-Yeung Shum. Natural image colorization. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pp. 309–320, 2007.

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.

Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

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.

Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE international conference on computer vision*, pp. 5898–5906, 2017.

David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3637–3645, 2018.

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

Kaiyue Pang, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10347–10355, 2020.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

Qi Qian, Yuanhong Xu, Juhua Hu, Hao Li, and Rong Jin. Unsupervised visual representation learning by online constrained k-means. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16640–16649, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Antoine Fleurentin, Georgios Exarchakis, et al. Dissecting self-supervised learning methods for surgical computer vision. *arXiv preprint arXiv:2207.00449*, 2022.

Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 762–771, 2018.

Pierre H Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, et al. Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6148–6157, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems*, 29, 2016.

Kihyuk Sohn and Honglak Lee. Learning invariant representations with local transformations. *arXiv preprint arXiv:1206.6418*, 2012.

Igor Susmelj, Matthias Heller, Philipp Wirth, Jeremy Prescott, and Malte Ebner et al. Lightly. *GitHub. Note: https://github.com/lightly-ai/lightly*, 2020.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*, pp. 270–279. Springer, 2018.

Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14431–14440, 2022.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv preprint arXiv:2301.03580*, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pp. 776–794. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020b.

Yonglong Tian, Olivier J Henaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10063–10074, 2021a.

Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020c.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021b.

Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? *arXiv preprint arXiv:2201.05119*, 2022.

Trieu H Trinh, Minh-Thang Luong, and Quoc V Le. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*, 2019.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pp. 268–285. Springer, 2020.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, 2008.

Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. Self-supervised learning by estimating twin class distributions. *arXiv preprint arXiv:2110.07402*, 2021a.

Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021b.

Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, Heli Qi, Zhen Wu, Yu-Feng Li, Satoshi Nakamura, Wei Ye, Marios Savvides, Bhiksha Raj, Takahiro Shinozaki, Bernt Schiele, Jindong Wang, Xing Xie, and Yue Zhang. Usb: A unified semi-supervised learning benchmark for classification. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022a. doi: 10.48550/arxiv.2208.07204. URL `https://arxiv.org/abs/2208.07204`.

Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Mingming Gong, and Tongliang Liu. Exploring set similarity for dense self-supervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16590–16599, 2022b.

Zhaoqing Wang, Ziyu Chen, Yaqian Li, Yandong Guo, Jun Yu, Mingming Gong, and Tongliang Liu. Mosaic representation learning for self-supervised visual pre-training. In *International Conference on Learning Representations*, 2023.

Zhirong Wu, Alexei A Efros, and Stella X Yu. Improving generalization via scalable neighborhood component analysis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 685–701, 2018a.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018b.

Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34:28864–28876, 2021a.

Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *International Journal of Computer Vision*, pp. 1–20, 2022a.

Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021b.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021c.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022b.

Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Bag of instances aggregation boosts self-supervised distillation. In *International Conference on Learning Representations*, 2021.

Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. *Advances in Neural Information Processing Systems*, 17, 2004.

Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6721–6729, 2017.

Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5147–5156, 2016.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4471–4480, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

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

Chaoning Zhang, Kang Zhang, Trung X Pham, Axi Niu, Zhinan Qiao, Chang D Yoo, and In So Kweon. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14441–14450, 2022a.

Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2547–2555, 2019.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pp. 649–666. Springer, 2016.

Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.

Shaofeng Zhang, Lyn Qiu, Feng Zhu, Junchi Yan, Hengrui Zhang, Rui Zhao, Hongyang Li, and Xiaokang Yang. Align representations with base: A new approach to self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16600–16609, 2022b.

Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16580–16589, 2022c.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. *Advances in Neural Information Processing Systems*, 34:2543–2555, 2021.

Yuanyi Zhong, Haoran Tang, Junkun Chen, Jian Peng, and Yu-Xiong Wang. Is self-supervised contrastive learning more robust than supervised learning? In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6002–6012, 2019.

Andrew Zisserman. Self-supervised learning. 2018. URL `https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf`.

Appendix for: Know Your Self-supervised Learning: A Survey on Image-based Discriminative Training

The appendix contains the following information:

- Section A provides the list of abbreviations for the SSL frameworks covered in the main text.
- Table 8 provides metadata for the SSL frameworks such as the primary affiliation, publication date, and source code availability.
- Tables (9 to 13) contains the benchmarking results on ImageNet-1k for clustering-based (Table 9), contrastive-learning-based (Table 10 and Table 11), distillation-based (Table 12), and information-maximization-based (Table 13) SSL frameworks.

The appendix contains additional information about SSL frameworks covered in the main text. list of abbreviations for SSL frameworks covered in the main text

A List of abbreviations

BERT	Bidirectional Encoder representations from transformers
GPT	Generative Pre-trained Transformer
SCAN	Semantic clustering by adopting nearest neighbors
Swav	Swapping assignments between multiple views of the same image
PIRL	Pretext-invariant representation learning
SimCLR	A simple framework for contrastive learning
NPID	Non-parametric instance discrimination
PCL	Prototypical contrastive learning
PIC	Parametric instance classification
DCL	Debiased contrastive learning
InterCLR	inter-image contrastive learning
G-SimCLR	Self-Supervised contrastive learning with guided projection
MocHi	Mixing of contrastive hard negatives
Note:	Mo of MocHi has nothing to do with Mo of MoCo
DenseCL	Dense contrastive learning
PixPro	Pixel-Level consistency propagation
CLSA	Contrastive learning with stronger augmentations
NNCLR	Nearest-neighbor contrastive learning of visual representations
SimMoCo	Simplified MoCo
SimCo	Simplified MoCo without momentum
UniVIP	A Unified framework for self-supervised visual pre-training
SimSiam	Simple Siamese representation learning networks
SEED	Self-supervised distillation for visual representation
DisCo	Distilled contrastive learning
DINO	Self-distillation with no labels
MSF	Mean shift for self-supervised learning
ORL	Object-level representation learning
BINGO	Bag of instances aggregation
WMSE	Whitening mean squared error
VicReg	Variance-invariance-covariance regularization
TWIST	Twin Class Distribution Estimation
TLDR	Twin learning for dimensionality reduction
ARB	Align representations with base
MosRep	Mosaic representation learning framework
ReSSL	Relational self-supervised learning
ReLIC	Representation learning via invariant causal mechanisms
SeLa	Self labeling
ODC	Online deep clustering
InstDist	Instance discrimination
CPC	Contrastive predictive coding
CMC	Contrastive multi-view coding
MoCo	Momentum contrast
AdCo	Adversarial contrast
BYOL	Build your-own latent
DIM	Deep InfoMax
AMDIM	Augmented multi scale DIM
MoBY	MoCo + BYOL
DnC	Divide and contrast
ReSSL	Relational self-supervised learning
UniGrad	Unified gradient framework
OBoW	Online Bag-of-visual-words
DirectPred	Direct linear predictor
CoKe	Constrained K-means
SMoG	Synchronous momentum grouping
CEB	Conditional entropy bottleneck
SEM	Simplicial embeddings

Table 8: Publication information as well as implementation details for the SSL frameworks covered in this survey.

SSL framework	Primary affiliation	Publication date	Experiments on on ImageNet 1K	Downstream experiments	Official implementation	Trained models
Deep Cluster	Facebook AI Research	Mar 2019	Yes	Yes	Available	Available
Local Aggregation	Stanford University	Apr 2019	Yes	Yes	Available	Not available
Deeper Cluster	Facebook AI Research	Aug 2019	Yes	Yes	Available	Available
SeLa	University of Oxford	Nov 2019	Yes	Yes	Available	Available
SCAN	KU Leuven	Jul 2020	Yes	No	Available	Available
Deep Cluster-v2	Facebook AI Research	Jun 2020	Yes	Yes	Available	Available
SeLa-v2	Facebook AI Research	Jun 2020	Yes	Yes	Available	Available
Swav	Facebook AI Research	Jun 2020	Yes	Yes	Available	Available
ODC	SenseTime	Jun 2020	Yes	Yes	Available	Not available
CoKe	Alibaba	May 2021	Yes	Yes	Available	Available
Self-Classifier	IBM Research	Jul 2022	Yes	Yes	Available	Available
InstDist (NPID)	Chinese Univ. of Hong Kong	May 2018	Yes	Yes	Available	Available
CPC	DeepMind	Jul 2018	Yes	No	Not available	Not available
DIM	Microsoft Research	Aug 2018	No	No	Available	Not available
CPC-v2	DeepMind	May 2019	Yes	Yes	Not available	Not available
AMDIM	Microsoft Research	Jun 2019	Yes	No	Available	Available
CMC	MIT	Jun 2019	Yes	Yes	Available	Available
MoCo	Facebook AI Research	Nov 2019	Yes	Yes	Available	Available
PIRL	Facebook AI Research	Dec 2019	Yes	Yes	Not available	Not available
SimCLR	Google Research	Feb 2020	Yes	Yes	Available	Available
MoCo-v2	Facebook AI Research	Mar 2020	Yes	Yes	Available	Available
InfoMin	MIT	May 2020	Yes	Yes	Available	Available
SimCLR-v2	Google Research	Jun 2020	Yes	No	Available	Available
PCL & PCLv2	Salesforce Research	Jun 2020	Yes	Yes	Available	Available
PIC	Microsoft Research	Jun 2020	Yes	Yes	Available	Available
DCL	MIT	Jul 2020	No	Yes	Available	Available
InterCLR	Nanyang Technological Uni.	Aug 2020	Yes	No	Available	Not available
G-SimCLR	Walmart Labs	Sep 2020	No	No	Available	Available
MocHi	Naver Labs EU	Oct 2020	Yes	Yes	Not available	Available
ReLIC	DeepMind	Oct 2020	Yes	Yes	Not available	Not available
AdCo	Peking University	Nov 2020	Yes	Yes	Available	Available
DenseCL	The University of Adelaide	Nov 2020	No	Yes	Available	Available
PixPro	Microsoft Research	Nov 2020	Yes	Yes	Available	Available
MoCo-v3	Facebook AI Research	Apr 2021	Yes	Yes	Available	Available
CLSA	Purdue University	Apr 2021	Yes	Yes	Available	Available
NNCLR	Google Research	Apr 2021	Yes	Yes	Not available	Not available
MoBY	Microsoft Research	May 2021	Yes	Yes	Available	Available
DNC	DeepMind	May 2021	Yes	Yes	Not available	Not available
ReSSL	SenseTime	Jul 2021	Yes	No	Available	Available
UniGrad	SenseTime	Dec 2021	Yes	No	Available	Available
ReLIC-v2	DeepMind	Jan 2022	Yes	Yes	Not available	Not available
SimCo	KAIST	Mar 2022	No	No	Available	Not available
SimMoCo	KAIST	Mar 2022	No	No	Available	Not available
UniVIP	University of Chinese AoS	Mar 2022	Yes	Yes	Not available	Not available
SMoG	Huawei	Jul 2022	Yes	Yes	Not available	Not available
MosRep	University of Sydney	Feb 2023	Yes	Yes	Available	Available
BYOL	DeepMind	Jun 2020	Yes	Yes	Available	Available
SimSiam	Facebook AI Research	Aug 2020	Yes	Yes	Available	Available
OBoW	Valeo.ai	Dec 2020	Yes	Yes	Available	Available
SEED	Microsoft Research	Jan 2021	Yes	Yes	Not available	Not available
DirectPred	Facebook AI Research	Feb 2021	Yes	Yes	Available	Not available
DisCO	Tencent	Apr 2021	Yes	Yes	Available	Available
DINO	Facebook AI Research	Apr 2021	Yes	Yes	Available	Available
MSF	University of Maryland	May 2021	Yes	No	Available	Available
ORL	Nanyang Technological University	Dec 2021	Yes	Yes	Available	Available
CEB	Google Research	Dec 2021	Yes	Yes	Available	Available
BINGO	Huawei	Mar 2022	Yes	Yes	Available	Not available
SEM	MILA	Apr 2022	Yes	Yes	Not available	Not available
WMSE	University of Trento	Jul 2020	Yes	No	Available	Not available
Barlow Twins	Facebook AI Research	Mar 2021	Yes	Yes	Available	Available
VicReg	Facebook AI Research	May 2021	Yes	Yes	Available	Available
TWIST	Tsinghua University	Oct 2021	Yes	Yes	Available	Available
TLDR	Naver Labs EU	Oct 2021	Yes	Yes	Available	Not available
ARB	Shanghai Jiao Tong University	Nov 2021	Yes	No	Not available	Not available

Table 9: ImageNet-1k benchmarks for **clustering**-based SSL frameworks. If the evaluation protocol deviates from commonly used methods for linear probing or KNN-based evaluation described in the main text, we provide additional notes in the rightmost column. "SSL epochs" denotes the number of epochs for the SSL training. Top-5 accuracy is empty for majority of the frameworks since that information is not available in their respective papers.

SSL framework	Backbone network	SSL epochs	Benchmarking method	Top-1 accuracy	Top-5 accuracy	Additional notes
Deep Cluster	AlexNet	500	Linear	39.8	-	Used conv4 output
LA	AlexNet	90	Linear	42.4	-	-
LA	VGG-16	90	Linear	57.6	-	-
LA	ResNet-18	90	Linear	52.8	-	-
LA	ResNet-50	90	Linear	60.2	-	-
LA	AlexNet	90	KNN	40.3	-	k=200
LA	VGG-16	90	KNN	50.0	-	k=200
LA	ResNet-18	90	KNN	45.6	-	k=200
LA	ResNet-50	90	KNN	50.1	-	k=200
Deeper Cluster	VGG-16	90	Linear	48.4	-	-
SeLA	ResNet-50	90	Linear	61.5	84.0	-
SCAN	ResNet-50	90	Linear	39.9	60.0	Unsupervised evaluation
Deep Cluster-v2	ResNet-50	400	Linear	70.2	-	-
Deep Cluster-v2	ResNet-50	400	Linear	74.3	-	2x160 + 4x96 crops
SeLA-v2	ResNet-50	400	Linear	67.2	-	-
SeLA-v2	ResNet-50	400	Linear	71.8	-	2x160 + 4x96 crops
Swav	ResNet-50	100	Linear	72.1	-	-
Swav	ResNet-50	200	Linear	73.9	-	-
Swav	ResNet-50	400	Linear	74.6	-	-
Swav	ResNet-50	800	Linear	75.3	-	-
Swav	ResNet-50	800	KNN	65.7	-	k=20
Swav	ResNet-50	800	KNN	62.7	-	k=200
ODC	AlexNet	100	Linear	41.4	-	-
ODC	ResNet-50	100	Linear	57.6	-	-
CoKe	ResNet-50	800	Linear	71.4	-	-
CoKe	ResNet-50	800	Linear	76.4	-	8 views
Self-C.	ResNet-50	100	K-means	37.3	-	Unsupervised evaluation
Self-C.	ResNet-50	200	K-means	39.4	-	Unsupervised evaluation
Self-C.	ResNet-50	400	K-means	40.2	-	Unsupervised evaluation
Self-C.	ResNet-50	800	K-means	41.1	-	Unsupervised evaluation
Self-C.	ResNet-50	100	Linear	72.4	-	-
Self-C.	ResNet-50	200	Linear	73.5	-	-
Self-C.	ResNet-50	400	Linear	74.2	-	-
Self-C.	ResNet-50	800	Linear	74.1	-	-

Table 10: ImageNet-1k benchmarks for **contrastive**-based SSL frameworks. If the evaluation protocol deviates from commonly used methods for linear probing or KNN-based evaluation described in the main text, we provide additional notes in the rightmost column. "SSL epochs" denotes the number of epochs for the SSL training.

SSL framework	Backbone network	SSL epochs	Benchmarking method	Top-1 accuracy	Top-5 accuracy	Additional notes
InstDist (NPID)	AlexNet	200	Linear	35.6	-	-
InstDist (NPID)	VGG-16	200	Linear	39.2	-	-
InstDist (NPID)	ResNet-18	200	Linear	44.5	-	-
InstDist (NPID)	ResNet-50	200	Linear	54.0	-	-
InstDist (NPID)	AlexNet	200	KNN	31.3	-	k=200
InstDist (NPID)	VGG-16	200	KNN	33.9	-	k=200
InstDist (NPID)	ResNet-18	200	KNN	41.0	-	k=200
InstDist (NPID)	ResNet-50	200	KNN	46.5	-	k=200
CPC	ResNet-v2 101	130	Linear	48.7	73.6	-
DIM	-	-	-	-	-	No benchmarks on ImageNet-1k
CPCv2	ResNet-50	200	Linear	61.8	85.3	-
CPCv2	ResNet-161	200	Linear	71.5	90.1	-
AMDIM	ResNet-50	150	Linear	63.5	-	Small AMDIM model
AMDIM	ResNet-50	150	Linear	68.1	-	Large AMDIM model
CMC	ResNet-50	200	Linear	66.2	87.0	-
CMC	ResNet-50 (×2)	200	Linear	70.6	89.7	-
MoCo	ResNet-50	200	Linear	60.6	-	-
MoCo	ResNeXt-50	200	Linear	63.9	-	-
MoCo	ResNet50 (×2)	200	Linear	65.4	-	-
MoCo	ResNet50 (×4)	200	Linear	68.6	-	-
PIRL	ResNet-50	800	Linear	63.6	-	-
SimCLR	ResNet-50	100	Linear	63.6	89.0	-
SimCLR	ResNet-50 (×2)	100	Linear	74.2	92.0	-
SimCLR	ResNet-50 (×4)	100	Linear	76.5	93.2	-
MoCo-v2	ResNet-50	200	Linear	67.5	-	-
MoCo-v2	ResNet-50	800	Linear	71.1	-	-
PCL	ResNet-50	200	Linear	61.5	-	-
PCL-v2	ResNet-50	200	Linear	67.6	-	-
PIC	ResNet-50	200	Linear	70.8	90.0	-
DCL	-	-	-	-	-	No benchmarks on ImageNet-1k
InterCLR	ResNet-50	200	Linear	68.0	-	with MoCo-v2
InterCLR	ResNet-50	200	Linear	73.5	-	with BYOL
InterCLR	ResNet-50	1000	Linear	74.5	-	with BYOL
G-SimCLR	-	-	-	-	-	No benchmarks on ImageNet-1k
MocHi	ResNet-50	800	Linear	70.6	-	with MoCo-v2
ReLIC	ResNet-50	800	Linear	74.8	92.2	-
AdCo	ResNet-50	200	Linear	68.6	-	-
AdCo	ResNet-50	200	Linear	73.2	-	Multi-crop
AdCo	ResNet-50	800	Linear	72.8	-	-
AdCo	ResNet-50	800	Linear	75.7	-	Multi-crop
DenseCL	-	-	-	-	-	No benchmarks on ImageNet-1k
PixPro	ResNet-50	100	Linear	66.3	-	with SimCLR

Table 11: ImageNet-1k benchmarks for **contrastive**-based SSL frameworks. If the evaluation protocol deviates from commonly used methods for linear probing or KNN-based evaluation described in the main text, we provide additional notes in the rightmost column. "SSL epochs" denotes the number of epochs for the SSL training. This table is a continuation of Table 2.

SSL framework	Backbone network	SSL epochs	Benchmarking method	Top-1 accuracy	Top-5 accuracy	Additional notes
MoCo-v3	ResNet-50	800	Linear	73.8	-	-
MoCo-v3	ViT-B	300	Linear	76.7	-	-
MoCo-v3	ViT-L	300	Linear	77.6	-	-
MoCo-v3	ViT-H	300	Linear	78.1	-	-
MoCo-v3	ViT-BN-H	300	Linear	79.1	-	-
MoCo-v3	ViT-BN-L/7	300	Linear	81.0	-	-
CLSA	ResNet-50	200	Linear	69.4	-	-
CLSA	ResNet-50	200	Linear	73.3	-	Multi-crop
CLSA	ResNet-50	800	Linear	72.2	-	-
CLSA	ResNet-50	800	Linear	76.2	-	Multi-crop
NNCLR	ResNet-50	1000	Linear	75.4.	92.3	-
NNCLR	ResNet-50	1000	Linear	75.6	92.4	Multi-crop
MoBY	DeiT-S	300	Linear	72.8	-	-
MoBY	Swin-T	100	Linear	70.9	-	-
MoBY	Swin-T	100	Linear	75.0	-	-
DNC	ResNet-50	1000	Linear	74.5	-	-
DNC	ResNet-50	1000	Linear	75.8	-	-
ReSSL	ResNet-50	200	Linear	69.9	-	-
ReSSL	ResNet-50	200	Linear	73.8	-	4 crops
ReSSL	ResNet-50	200	Linear	74.7	-	8 crops
UniGrad	ResNet-50	100	Linear	70.3	-	-
UniGrad	ResNet-50	100	Linear	71.2	-	with CutMix
UniGrad	ResNet-50	100	Linear	71.7	-	with multi-crop
UniGrad	ResNet-50	100	Linear	72.3	-	with CutMix + multi-crop
UniGrad	ResNet-50	800	Linear	75.5	-	with CutMix + multi-crop
ReLIC-v2	ResNet-50	1000	Linear	77.1	-	-
SimCo	-	-	-	-	-	No benchmarks on ImageNet-1k
SimMoCo	-	-	-	-	-	No benchmarks on ImageNet-1k
UniVIP	ResNet-50	300	Linear	74.2	-	-
SMoG	ResNet-50	400	Linear	73.6	91.3	-
SMoG	ResNet-50	800	Linear	74.5	91.9	-
SMoG	ResNet-50	400	Linear	76.4	93.1	with multi-crop
MosRep	ResNet-50	200	Linear	76.2	-	-

Table 12: ImageNet-1k benchmarks for **distillation**-based SSL frameworks. If the evaluation protocol deviates from commonly used methods for linear probing or KNN-based evaluation described in the main text, we provide additional notes in the rightmost column. "SSL epochs" denotes the number of epochs for the SSL training.

SSL framework	Backbone network	SSL epochs	Benchmarking method	Top-1 accuracy	Top-5 accuracy	Additional notes
BYOL	ResNet-50	1000	Linear	74.3	91.6	-
BYOL	ResNet-50 (×2)	1000	Linear	77.4	93.6	-
BYOL	ResNet-50 (×4)	1000	Linear	78.6	94.2	-
BYOL	ResNet-200 (×2)	1000	Linear	79.6	94.8	-
SimSiam	ResNet-50	100	Linear	68.1	-	-
SimSiam	ResNet-50	200	Linear	70.0	-	-
SimSiam	ResNet-50	400	Linear	70.8	-	-
SimSiam	ResNet-50	800	Linear	71.3	-	-
OBoW	ResNet-50	200	Linear	73.8	-	-
DirectPred	ResNet-50	60	Linear	64.4	85.8	-
DirectPred	ResNet-50	300	Linear	72.4	91.0	-
DINO	ResNet-50	300	Linear	75.3	-	-
DINO	ViT-S	300	Linear	77.0	-	-
DINO	ViT-B/16	300	Linear	78.2	-	-
DINO	ViT-S/8	300	Linear	79.7	-	-
DINO	ViT-B/8	300	Linear	80.1	-	-
DINO	ResNet-50	300	KNN	67.5	-	k=20
DINO	ViT-S	300	KNN	74.5	-	k=20
DINO	ViT-B/16	300	KNN	76.1	-	k=20
DINO	ViT-S/8	300	KNN	78.3	-	k=20
DINO	ViT-B/8	300	KNN	77.4	-	k=20
MSF	ResNet-50	200	Linear	72.4	-	-
MSF	ResNet-50	200	KNN	62.0	-	k=1
MSF	ResNet-50	200	KNN	65.2	-	k=20
ORL	ResNet-50	800	Linear	59.0	-	SSL training with COCO
ORL	ResNet-50	800	Linear	60.7	-	SSL training with COCO+
CEB	ResNet-50	1000	Linear	75.6	-	with BYOL
CEB	ResNet-50 (×2)	1000	Linear	78.8	-	with BYOL
CEB	ResNet-50	1000	Linear	71.0	-	with SimCLR
CEB	ResNet-50 (×2)	1000	Linear	75.0	-	with SimCLR
BINGO	ResNet-18	200	Linear	64.0	85.7	Distilled from ResNet-50
BINGO	ResNet-34	200	Linear	66.1	87.2	Distilled from ResNet-50
SEM	ResNet-50	200	Linear	74.1	-	with BYOL

Table 13: ImageNet-1k benchmarks for **information-maximization**-based SSL frameworks. If the evaluation protocol deviates from commonly used methods for linear probing or KNN-based evaluation described in the main text, we provide additional notes in the rightmost column. "SSL epochs" denotes the number of epochs for the SSL training.

SSL framework	Backbone network	SSL epochs	Benchmarking method	Top-1 accuracy	Top-5 accuracy	Additional notes
WMSE	ResNet-50	100	Linear	69.4	-	$d = 4$, corresponding to 6 positive pairs
WMSE	ResNet-50	400	Linear	72.5	-	$d = 4$, corresponding to 6 positive pairs
Barlow Twins	ResNet-50	1000	Linear	73.2	91.0	-
VicReg	ResNet-50	1000	Linear	73.2	91.1	-
TWIST	ResNet-50	800	Linear	72.6	91.0	-
TWIST	ResNet-50	800	Linear	75.5	92.5	Multi-crop
TWIST	Wide ResNet	300	Linear	77.7	93.9	-
TWIST	Deit-S/16	300	Linear	76.3	92.7	-
TWIST	ViT-B/16	300	Linear	78.4	93.8	-
ARB	ResNet-50	100	Linear	68.2	88.9	-