Artificial Intelligence Review (2026) 59:112
https://doi.org/10.1007/s10462-026-11506-9

A survey on design choices for self-supervised learning in computer vision

Ladyna Wittscher[1]

Received: 13 July 2025 / Accepted: 17 January 2026

Abstract
Self-supervised learning is a successful strategy to overcome data scarcity and improve robustness in computer vision by adding a pretext task that can exploit inherent data relationships as supervision signals during pretraining. However, the combination of pretraining and downstream training renders model design more complex, as additional design choices are required. This paper analyses the effects of such design choices specific to self-supervised learning on model performance and robustness. How does the pretext task influence the downstream task and how to design an ideal and generalizable pretext task? Which properties of the pretraining dataset are favorable and how similar should the pretext and downstream dataset ideally be? To address these questions, a comprehensive survey has been conducted, encompassing the results of diverse models and publications with different design choices. The results demonstrate the advantages of in-domain pretraining and the importance of aligning all design choices in order to ensure optimal results. Furthermore, the characteristic differences between predictive, contrastive and generative self-supervised learning and the design choices which are crucial for each of these learning paradigms are analyzed in detail.

Keywords Self-supervised learning · Design choice · Contrastive learning · Masked image modeling · Computer vision · Hyperparameter

1 Introduction

Self-supervised learning (SSL) can reduce the dependency of artificial neural networks (ANN) on large annotated datasets, which present a major bottleneck for deep learning due to their high cost and partial unavailability (Pal and Balasubramanian 2018; Sammani et al. 2023). At the same time, SSL can improve model robustness and generalization compared to supervised learning (SL) in computer vision (CV) (Hendrycks et al. 2019; Liu et al.

✉ Ladyna Wittscher
Ladyna.Wittscher@uni-jena.de

[1] Chair of Economic and Social Statistics, Friedrich-Schiller-University, Fürstengraben 1, 07743 Jena, Germany

Published online: 24 February 2026

2021a; Navarro et al. 2021; Zhao et al. 2024; Zhong et al. 2022) due to the higher "per-class feature uniformity" (Zhong et al. 2022), the larger number and diversity of pretraining images (Navarro et al. 2021), learning "label-irrelevant-but-transferable features" (Liu et al. 2021a) and improved feature extraction and data-efficiency (Zhao et al. 2024). SSL is also less prone to overfitting and generates more information per sample (Ericsson et al. 2021; Mao 2020).

SSL should not be confused with semi-supervised learning, which makes use of a small amount of labeled data combined with a large amount of unlabeled data (Van Engelen and Hoos 2020). Semi-supervised training can be sequential with a model being trained on the labeled data to predict unlabeled data and a second step where the full dataset is used for training (Rani et al. 2023). Recent semi-supervised models train jointly with one objective and two loss terms, one optimized for "labeled-set accuracy and the other favoring label-consistency or label-smoothness" (Huang et al. 2024a). However, SSL generates its own pseudo-labels and supervisory signals from the data itself by adding an upstream pretraining (or pretext) task (Gui et al. 2024; Huang et al. 2024a). Semi-supervised learning is mostly understood as a hybrid of supervised and unsupervised learning (Van Engelen and Hoos 2020). SSL can be classified as a sub-type of unsupervised learning as both strive to exploit data-inherent structures of large unlabeled datasets without human annotations (Chen et al. 2022d; Rani et al. 2023). Consequently, both approaches differ fundamentally.

Self-supervised pretext tasks aim to learn feature representations useful to the downstream task (Albelwi 2022), thereby improving the overall performance. The operating principle of SSL is illustrated in Fig. 1. SSL pretraining can either be executed with the same dataset as the downstream training (self-pretraining), with a distinct dataset from the same domain (in-domain pretraining), or with pretraining data from a completely different context (out-of-domain pretraining) (Jonske et al. 2025; Krishna et al. 2023; Zhang et al. 2024b). After the pretext task, the learned representations are transferred to the downstream training objective via knowledge transfer (KT), predominantly using either fine-tuning (FT) or linear probing (LP) (Ozbulak et al. 2023). While the pretext task is always self-supervised, the downstream task can be either unsupervised or supervised. Convolutional Neural Networks (CNN) and Vision Transformers (ViT) are the most important network architectures for discriminative models in CV and most research focuses on them. Nevertheless, SSL has been applied to other architectures, such as Capsule Networks (El-Shimy et al. 2025;

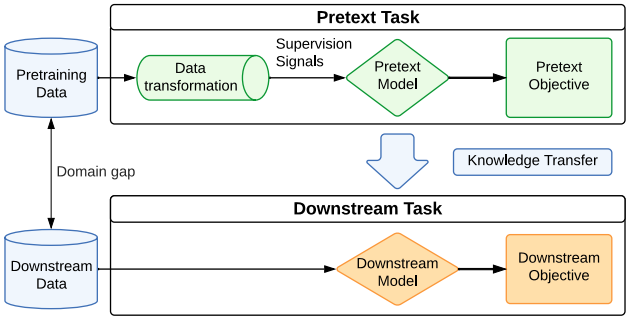

Fig. 1 The pretext model is trained using a pretext objective, typically after applying transformations to the pretraining data (e.g. masking or cropping). The resulting learned representation is then transferred to the downstream model with its own downstream objective. The domain gap depends on how much the pretraining and the downstream dataset differ

Everett et al. 2025; Sun et al. 2021; Tran et al. 2022) or Graph Neural Networks (Liu et al. 2022b; Xie et al. 2023c; Wei et al. 2024b). In principle, SSL is architecture-agnostic (Li et al. 2023d).

The performance of SSL models has approached and exceeded its SL counterparts (Ericsson et al. 2021; Goyal et al. 2021; Liu et al. 2023e; Sammani et al. 2023; Tomasev et al. 2022; Yang et al. 2025). Especially self-supervised Large-Language-Models (LLMs) like BERT and GPT revolutionize Natural Language Processing (NLP) and excel across multiple benchmarks due to the large volumes of text data available for training, while these models have also inspired adaptation and refinements for self-supervised CV (Jaiswal et al. 2020; Ozbulak et al. 2023). Still, images are highly redundant in contrast to words, which decreases generalization of research results obtained with LLM and makes it necessary to design specific pretext tasks for CV (Xie et al. 2023d). SSL can follow different learning paradigms: While generative approaches dominate within self-supervised NLP, for CV contrastive and generative approaches are equally important (Liu et al. 2021b), while hybrid approaches have gained more recognition recently (Shi et al. 2023b; Zong et al. 2024). Predictive approaches have been important for the development of visual SSL and are therefore included in this paper.

Due to the combination of pretext and downstream task, SSL requires additional design choices compared to standard SL. Design choices refer to decisions necessary during the development of a model which have a significant impact on the downstream performance. Design choices specific to SSL target the pretext task design and its influence on the downstream task, the relation of pretext and downstream data, KT from the pretext to the downstream task as well as design choices specific to one specific SSL learning paradigm. In contrast to hyperparameters, design choices are not limited to specific parameter values, but take more complex considerations or certain groups of hyperparameters into account. Design choices can influence the effectiveness of the learning process itself as they determine how well the model can learn from unlabeled data and generalize to various applications. Consequently, they are crucial for the model's robustness, e.g. regarding overfitting and adversarial attacks (Çağatan et al. 2025; Chander et al. 2025; Duesterwald et al. 2019; Markus et al. 2021). A more comprehensive understanding of the complex interplay between various design choices is necessary to facilitate further development of SSL methods. Despite the importance of SSL for CV and the numerous studies that propose novel SSL methods, few have systematically studied design choices. This is the first survey that in detail analyzes the impact of pretext task design, pretraining data, network architecture and size, KT, as well as learning-paradigm-specific design choices on the performance of SSL. We focus is on discriminative CV, where SSL is of significant importance. Consequently, this research makes an important contribution to the understanding of the specific behavior of SSL by giving and overview of the different design choices and their interdependence.

Section 2 discusses the concept of design choices in detail, Sect. 3 presents the methodology of this survey, and Sect. 4 analyzes related work. The different SSL learning paradigms and their characteristics are discussed in Sect. 5. The impact of pretext task design, such as its complexity, representation geometry, the combination of multiple pretext tasks, and the general influence of the pretext on the downstream task is analyzed in Sect. 6. Section 7 explores the implications of the network size. Section 8 evaluates the influence of the pretext data, including its characteristics, quality and size, the focus on ImageNet in research, as well as in-domain (ID) pretraining and self-pretraining. Sections 9–11 focuses on design

choices specific to one SSL paradigm. The impact of KT, including the choice between LP and FT, their performance, as well as the question which strategy is better at bridging domain gaps, is addressed in Sect. 12. Finally, Sect. 13 concludes the key findings and gives guidelines for practitioners.

2 Design choices

Design choices refer to modeling decisions that determine the structure, learning behavior and optimization constraints of ANNs. Like hyperparamters, design choices cannot be directly learned from the data and must be specified prior to training as they determine the conditions of training itself. $\mathcal{S}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$ describes the training dataset with n samples, $\theta \in \Theta$ the trainable model parameters and $d \in \mathcal{D}$ the design decicions. In SSL, during pretraining, the target y_i is generated by the pretext task also specified by *d*. The training algorithm (e.g. SGD) induced by *d* provides parameters

$$\hat{\theta}(d) = \underset{\theta \in \Theta(d)}{\arg\min}\, \mathcal{L}_{\text{train}}(\theta; d),$$

as an approximate solution to the empirical minimization problem. Both the trainable parameter domain $\Theta(d)$ and the training loss $\mathcal{L}_{\text{train}}$ consequently depend on the design choice. In contrast to hyperparameters, the concept of design choices is broader: it refers to groups of interrelated subdecisions (including classic hyperparameters) that should be considered jointly due to their mutually dependent nature. A design choice *d* consists of a set of *m* interrelated subdecisions

$$d = (z_1, \ldots, z_m), \qquad z_i \in \mathcal{Z}_i.$$

Not every combination of subdecisions is permissible. The set of valid design decisions is given by a restricted configuration set

$$\mathcal{D} \subseteq \prod_{i=1}^{m} \mathcal{Z}_i.$$

This restriction models the mutual influence of the subdecisions. It is possible to distinguish non-parameterizable subdecisions (categorical or binary choices), where Z_i is a finite set (e.g., learning paradigm $lp \in$ {contrastive, predictive, generative, hybrid}), from parameterizable subdecisions (continuous or integer-valued), where $Z_i \subset \mathbb{R}^k$ or $Z_i \subset \mathbb{Z}^k$ with specified bounds (e.g., augmentation probability $p \in [0, 1]$, encoder output dimension $e \in [128, 4096]$, number of negative samples $s \in \mathbb{N}$). Most SSL design choices are mixed: a categorical decision activates a set of conditional numeric parameters, yielding a hierarchical and conditional configuration space *D*. One example is the design of a suitable data augmentation strategy, which e.g. includes decisions which data augmentation techniques to apply (e.g. random horizontal flip and random cropping), how to combine them (e.g. the order of application), and which hyperparameters to use for the individual data augmenta-

tions (e.g. flip probability and crop size). Such decisions belong together and influence each other greatly, so they should not be optimized in isolation. Table 1 illustrates further examples.

The term design choice has been used before (see e.g. Arboretti et al. (2022a, 2022b); Li et al. (2021b); Mojvzivsek et al. (2025); Smith and Topin (2016); Villarraga and Daziano (2025)), but never analyzed systematically with a focus on SSL. Design choices often significantly influence the downstream performance, e.g. the low performance of SimCLR (Chen et al. 2020b) compared to VICReg (Bardes et al. 2021) can be compensated by simply adjusting some design choices (Garrido et al. 2022). They can influence the effectiveness of the learning process itself as they determine how well the model can learn from unlabeled data and generalize to various applications. Similar to hyperparameters, design choices can have a significant impact on robustness and explainability (Chander et al. 2025; Duesterwald et al. 2019; Markus et al. 2021). Still, design choices are often driven by incidental factors such as the developer's experience (Arboretti et al. 2022b), so a structured analysis can help to improve decision making (Arboretti et al. 2022a). If design choices and their impacts are not carefully studied, improvements reported in one paper cannot be replicated due to slightly modified design choices (Zhang et al. 2023c). Furthermore, they are often only studied on benchmarking data, which could lead to dataset-specific biases. Analyzing them on a wider range of datasets and domains can foster more generalizable conclusions.

The design choices considered in this paper cover all important steps in the development of SSL: data selection, design of the pretext task, the relationship between the pretext and the downstream task, the choice of a suitable network architecture, and the KT from pretext to downstream task. Figure 2 presents all discussed design choices as well as their role in SSL. This paper differentiates between two kinds of design choices: Those that are required for all SSL learning paradigms (Sects. 6–8 and 12) and those that are only relevant for one specific SSL learning paradigm (Sects. 9–11). An extensive literature review was conducted on recent developments in discriminative SSL for images and the insights regarding critical factors influencing SSL performance were extracted and organized according to distinct stages of the model development process. Subsequently, only those design choices were retained that have been consistently addressed across multiple studies and supported by substantial empirical evidence. Design choices or subdecisions that were only relevant for a few, very specific models or that were not specific to SSL were not included. Instead,

Table 1 Examples of design choices and corresponding subdecisions including the subdivision type and exemplary values. The table is not a complete list and is only intended to illustrate the concept

Design choice	Subdecision	Subdecision type	Exemplary values
Pretext task	Learning paradigm	Non-parameterizable (categorical)	predictive, contrastive, generative, hybrid
Pretext task	Multiple pretext tasks	Non-parameterizable (binary)	yes, no
Pretext task	Task	Non-parameterizable (categorical)	rotation, jigsaw, inpainting, exemplar
Pretext task	Rotation angle	Parameterizable (continuous)	$\alpha \in [0°, 360°]$
Pretext task	Rotation probability	Parameterizable (continuous)	$p \in [0, 1]$
Augmentation	Augmentation Strategy	Non-parameterizable (categorical)	Crop, Mask, Flip
Augmentation	Masking strategy	Non-Parameterizable (categorical)	Random, Block, Multi
Augmentation	Masking ratio	Parameterizable (continuous)	$r \in [0, 1]$
Augmentation	Patch size	Parameterizable (integer)	$t \in [1, 32]$

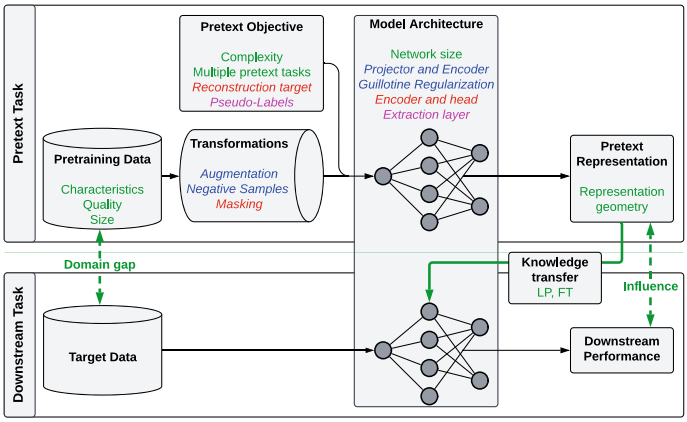

Fig. 2 Overview of important SSL design choices. Most choices (pretraining data characteristics, quality and size, domain gap, pretext complexity, multiple pretext tasks, network size, representation geometry, KT, influence of the pretext on the downstream task) are relevant to all paradigms (green). Other choices are particularly important for one paradigm (italic), such as pseudo-labels and extraction layers for predictive SSL (magenta). Augmentation, negative samples, Guillotine Regularization, projector and encoder architecture are specific to contrastive SSL (blue). Reconstruction target, masking, encoder and head architecture are important to generative SSL (red)

the focus is on those design choices that are relevant to the general functionality of SSL. Network size is included due to its tremendous impact both on the performance and on multiple design choices. Other classical hyperparameters, such as dropout, training epochs, optimization algorithm, loss function or batch normalization are not included in this survey because these considerations are less specific to SSL and a general optimization issue of ANN.

3 Methodology

The information source identification and search strategy follows the steps outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Moher et al. 2010), but was adapted. Papers rarely focus on design choices as their main topic but instead address related aspects or include relevant data e.g. in their ablation studies. Consequently, a simple keyword search is not sufficient. The strategy to identify relevant articles therefore consists of three pillars: Traditional keyword search in scientific databases, semantic search using AI, and cross-references in relevant publications to other publications. The keyword search was conducted using Google Scholar and Web of Science (core collection). To ensure that the publications focus on 'self-supervised' and 'images', these keywords were included in every search. Search was conducted both on design choices in general as well as on specific design choices. The exact keywords used can be found in Table 2. The first thirty hits were taken into account for each search. Consequently, 7680 articles were identified. In addition, ChatGPT and Consensus were used for semantic search, identifying 1386 articles. Basically, the same keywords were used, but very similar keywords, e.g. big and large or pretext data and pretext dataset, were excluded due to the semantic understanding of these models. In both cases, titles and snippets were

Table 2 Keywords used for database search sorted by topic. Every search additionally included 'self-supervised' and 'image' to ensure that the publications focused on these topics. Synonyms of keywords were included

Topic	Keywords
Design choices	design choice, model design, design, model configuration, hyperparameter
Related work	Review, survey, multi-modal, large pretraining, large model pretraining, large model, large architecture, large network, big model, big architecture, big network
Learning paradigms	learning paradigms, paradigms, frameworks, types, contrastive, generative, masked image modeling, MIM, predictive, hybrid
Pretext task design	pretext, pretext design, pretext complexity, pretext strength, pretext generalization, generalization, pretext downstream correlation, influence pretext downstream, pretext evaluation, pretext evaluation metric, representation geometry, representation dimensionality, multiple pretext tasks, multiple pretexts, pretext combination, combining pretext tasks
Network size	network size, architecture size, model size, size, network capacity, architecture capacity, model capacity, capacity, large networks, large architecture, large models, big networks, big models, big, large, small networks, small architecture, small model, small, model size, small architecture, small models
Pretext data	pretext data, pretext dataset, pretext dataset size, data size, dataset size, low-data regimes, large dataset, big dataset, small dataset, data quality, dataset size model size, dataset size, ImageNet pretraining, in-domain pretraining, domain gap, out-of-domain pretraining, in-domain, out-of-domain, self-pretraining
Design choices for predictive SSL	predictive models, predictive design, predictive difficulty, predictive distinguishable, predictive feature extraction, predictive feature extraction layer
Design choices for contrastive SSL	contrastive models, contrastive design, contrastive data augmentation, contrastive augmentation, data augmentation, augmentation, contrastive strength augmentation, augmentation strength, contrastive negative, contrastive negative samples, contrastive negative sampling, contrastive projector, contrastive guillotine regularization, guillotine regularization, contrastive encoder
Design choices for generative SSL	generative models, generative design, masked image modeling design, masking, MIM masking, MIM masking strategy, masking strategy, MIM masking ratio, masking ratio, MIM ratio, reconstruction target, generative encoder, MIM encoder, MIM head, generative head
Knowledge transfer	knowledge transfer, representation transfer, transfer learning, task transfer, finetuning, finetuning, linear probing, linear evaluation, domain gap

used to identify relevant publications. The following inclusion criteria were applied during screening: focus on self-supervised learning, focus on images/computer vision, focus on design choices or some aspects of design choices. Criteria of exclusion were duplicated articles. These criteria were met by 1216 articles.

In the eligibility phase, the full text was screened and the following additional criteria for inclusion were applied: The article discusses design choices or aspects of design choices in sufficient detail. The following additional criteria for exclusion were applied: The findings of the article only relate to a very specific method and are neither comparable nor generalizable, the article publishes no exact values of experimental results, the article does not publish all relevant hyperparameters and design choices. The last aspect is important to ensure that the results are comparable and to highlight the interdependence of different design choices. When scanning the text, references to other publications that could potentially be relevant were identified. If the title met the criteria of the screening phase, it was also examined according to the eligibility criteria. Consequently, 355 articles could be included in this survey. Stricter criteria were applied for the following tables and figures which focus on comparing specific design choices to achieve comparability, as SSL

is very sensitive to hyperparameter settings (Huang et al. 2024a): Table 15 focuses on results applying a fixed masking ratio, Table 17 focuses on contrastive models that self-pretrain on ImageNet-1k, Table 18 on generative and hybrid models that self-pretrain on ImageNet-1k and Table 19 on models pretrained with ImageNet applying downstream datasets for which Maximum Mean Discrepancy (MMD) (Chen et al. 2022b) results are available. Figure 7 solely includes results obtained on ResNet-50 with LP on ImageNet-1K, while Fig. 8 only includes results obtained on ViT models on ImageNet-1K.

4 Related work

Since 2020, numerous important review papers on SSL for CV were published, reflecting the rapid progress and growing relevance of SSL. Tables 3 and 4 provided an overview of important reviews (sorted by publication year and first author). Their characteristics, thematic focus and categorization of different SSL paradigms are discussed in this section. While only review papers focusing on images were included, many of these also discuss other data types: One third addresses images only, around 40% include video or multimodal methods, while text is included by a quarter and audio by less than a fifth. Graph, point clouds and code are rarely included. Earlier work focused more on the functionality of SSL, while more recent publications rather address specific applications. SSL is very popular for medical applications (see e.g. Choi (2024); Pani and Chawla (2025); Wang et al. (2023a); Zeng et al. (2024)), while few reviews focus on astronomy (Hayat et al. 2021), biology (Al Mamun et al. 2024), depth estimation (Dong et al. 2025) or remote sensing (Berg et al. 2022; Wang et al. 2022a). Most reviews focus on the application of SSL, either to a specific domain or to several. More than half of the articles focus on the functionality of SSL itself, often discussing in detail how methods work. One quarter of the publications in Table 3 and 4 either compare different SSL methods (e.g. Khan et al. (2022, 2025b); Ozbulak et al. (2023)) or SSL with other approaches, such as transfer learning (Zhao et al. 2024) or weakly- and semi-supervised learning (Qu et al. 2022). Shwartz Ziv and LeCun (2024) focus on developing a theoretical framework for SSL, they elaborate on the intersection of information theory, SSL, and deep neural networks. Few reviews discuss the implementation of SSL: Ericsson et al. (2022b) analyze data requirements, architecture choices, the transferability of the pretext features and how to choose the optimal pretext task. Huang et al. (2023) address the performance of different learning paradigms, the impact of KT and implementation guidelines. Zhang et al. (2023c) focus on data bias in pretext data, class-imbalance, overfitting in projector representations and the limits of data augmentation. Taherdoost (2024) discuss the importance of evaluation metrics, especially area under the curve (AUC). Yet, none of these studies systematically examines the underlying design choices or their interrelationships in depth in order to provide guidelines for practitioners on how SSL should be designed in order to achieve robust, generalizable and high-performance results. This article addresses that gap by analyzing insights and data from existing reviews and empirical studies with regard to design choices. While previous surveys often focus on categorizing existing SSL methods, discussing specific applications and comparing performance across benchmarks, they rarely examine the underlying design principles that determine the performance of different approaches. In contrast, this work systematically analyzes

Table 3 Overview of reviews addressing SSL with images with different main topics (Focus): functionality of SSL methods (M), application to one or several domains (A), development of a theoretical framework (Th), comparison of different methods (C), practical implementation (I) and research trends (Tr). Reviews divide SSL methods into different learning paradigms (Paradigms) and include different types of data (Data types). If a paradigm is mentioned but not discussed in detail, it is in brackets. The listed papers focus on different domains (Domains), some on SSL in general without a specific application (General), some address multiple domains in detail (Multiple), others focus on one single domain

Review article	Focus	Paradigms	Data types	Domain
Jaiswal et al. (2020)	M	Contrastive, (generative)	Image, video	General
Jing and Tian (2020)	M	Generation, context, free semantic label, cross modal	Image, video, audio, multimodal	General
Chowdhury et al. (2021)	M, A	Contrastive, adversarial, pixel to pixel, pixel to scalar	Image, video	Medicine
Hayat et al. (2021)	A	Contrastive	Image	Astronomy
Ohri and Kumar (2021)	M	Contrastive, clustering, reconstruction, generation, prediction	Image, multimodal	General
Xu (2021)	A	Generated, context, contrast	Image	Medicine
Albelwi (2022)	M	Contrastive, auxiliary pretext	Image	General
Berg et al. (2022)	M, A, C	(Non-)Contrastive, generative, predict	Image, video, multimodal	Remote sensing
Ericsson et al. (2022b)	M, C, I	Masked/transformation prediction, instance discrimination, clustering	Image, text, audio, graph	General
Huang et al. (2022)	A, Tr	Contrastive, prediction, self-distillation, clustering	Image, video	General
Khan et al. (2022)	M, C	Contrastive, (generative)	Image	General
Krishnan et al. (2022)	A	Contrastive, generative	Image, text, video, multimodal	Medicine
Kumar et al. (2022b)	M, A, Tr	Contrastive, generative	Image, video, text, audio, code, graphs, multimodal	Multiple
Qu et al. (2022)	A, C	Contrastive, generative, predictive	Image	Medicine
Shurrab and Duwairi (2022)	A	Contrastive, generative, predictive	Image, video	Medicine
Wang et al. (2022a)	M, A	Contrastive, generative, predictive	Image, video	Remote sensing
Zhang et al. (2023b)	M, A	(Contrastive), generative	Image, cloud points, graph, audio	Multiple
Huang et al. (2023)	A, I	Contrastive, generative, innate relationship, self-prediction	Image	Medicine
Liu et al. (2023e)	M	Contrastive, generative, adversarial	Image, text, audio, multimodal	General
Ozbulak et al. (2023)	M, C, Tr	Discriminative, generative	Image	General
Rani et al. (2023)	M, A	(Non-)Contrastive, auxiliary pretext	Image, video, text	Multiple

how the choice of data, pretext task, learning paradigm, architecture and KT interact and influence model performance.

4.1 Learning paradigms

Most reviews divide SSL methods into different learning paradigms. We discuss the different concepts of the reviews listed in Tables 3 and 4 to show which definitions overlap with

Table 4 Continuation of the overview of review papers for SSL with images including learning paradigms, data types and domains. They focus on methods (M), application (A), theory (Th), comparison (C), implementation (I) or trends (Tr)

Review article	Focus	Paradigms	Data types	Domain
Uelwer et al. (2025)	M, C	Contrastive, clustering, pretext task, information maximization, teacher-student	Image	General
Wang et al. (2023a)	A, Tr	Contrastive, (generative), predictive	Image, video, multimodal	Medicine
Zhang et al. (2023c)	A, I	Contrastive, generative, predictive, multi	Image	Medicine
Abdulrazzaq et al. (2024)	A	(Non-)Contrastive, generative, predictive	Image	Multiple
Al Mamun et al. (2024)	A	Contrastive, generative, predictive, hybrid	Image	Biology
Chen et al. (2024b)	M, A, Tr	Contrastive, generative, predictive, MIM	Image, video	Multiple
Choi (2024)	M, A	–	Image	Medicine
Ciocarlan et al. (2024)	A, C	Instance discrimination, MIM	Image, video	Multiple
Gui et al. (2024)	M, A, Tr	Contrastive, generative, context, contrastive-generative	Image, video, text, multimodal	Multiple
Liu et al. (2024c)	A, Tr	Contrastive, pseudo-label	Image, multimodal	Medicine
Rani et al. (2024)	A	Contrastive, generative, (predictive)	Image, multimodal	Medicine
Shwartz Ziv and LeCun (2024)	Th	Generative, joint-embedding	Image	General
Taherdoost (2024)	A, I	Contrastive, generative, predictive	Image, text	Medicine
VanBerlo et al. (2024)	A	Contrastive, (generative, predictive)	Image, video, multimodal	Medicine
Wang et al. (2024b)	M, C, Tr	Contrastive, generative, contrast.-generative	Image, video, text, multimodal	General
Zeng et al. (2024)	A	Contrastive, generative	Image	Medicine
Zhang et al. (2024d)	A	–	Image	General
Zhao et al. (2024)	C	Contrastive, generative, predictive	Image, video, text, audio, graph, multimodal	Multiple
Zong et al. (2024)	M, A	Instance discrimination, clustering, masked prediction, hybrid	Multimodal	Multiple
Dong et al. (2025)	A, C, Tr	–	Image	Depth
Hondru et al. (2025)	M, C	Contrastive/reconstruction MIM	Image, video, text, audio, multi	General
Khan et al. (2025b)	M, C, Tr	Contrastive, generative, clustering, knowledge distillation, hybrid	Image, multimodal	General
Khan et al. (2025a)	M, A	Contrastive	Multimodal	Multiple
Li et al. (2025)	A	–	Image	Medicine
Martinović et al. (2025)	A	Contrastive, generative	Image, multimodal	Medicine
Pani and Chawla (2025)	A, C	Contrastive, generative, predictive	Image	Medicine
Zhang et al. (2025b)	M, A, Tr	Contrastive, generative, context, clustering, graph, hybrid	Image, video, audio, text, graph	Multiple

our classification and which ones are distinct. The goal is to generate a better understanding regarding the character of the different paradigms. Hayat et al. (2021) and Khan et al. (2025b) only focus on contrastive approaches, while Choi (2024); Dong et al. (2025); Zhang et al. (2024d) do not distinguish between different paradigms. The majority, more than half of all review papers, distinguishes between contrastive and generative approaches. While contrastive approaches learn to distinguish between similar and dissimilar samples, generative approaches learn to reconstruct or generate data from its latent representation (Liu et al. 2024b). More than one quarter of all reviews additionally includes predictive learning, which is based on pseudo-labels and includes early SSL pretext tasks such as inpainting, rotation or jigsaw. These three paradigms are in accordance with the historical development of SSL (Wang et al. 2022a), but due to the decreasing importance of predictive SSL, many authors focus on contrastive and generative approaches only (Liu et al. 2021b; Kumar et al. 2022b; Ozbulak et al. 2023). Still, we include predictive approaches into our consideration due to their importance for the development of SSL. Like Al Mamun et al. (2024), we also consider hybrid approaches that have become increasingly important in recent years, due to the attempt to create synergies and take advantage of complementary strengths (Zong et al. 2024). The categorization of learning paradigms in this paper is discussed in detail in Sect. 5.

Chowdhury et al. (2021); Gui et al. (2024); Liu et al. (2023e); Wang et al. (2024b) include contrastive-generative learning or adversarial learning, which can also be considered as a type of generative or hybrid learning. While clustering-based approaches can be seen as a sub-type of CL (e.g. Jaiswal et al. (2020); Liu et al. (2023e), several reviews (Ericsson et al. 2022a; Huang et al. 2022; Khan et al. 2025b; Ohri and Kumar 2021; Uelwer et al. 2025; Zhang et al. 2025b; Zong et al. 2024) define it as an independent category. Discriminative learning (Ozbulak et al. 2023), joint-embedding (Shwartz-Ziv et al. 2022) and instance discrimination (Ciocarlan et al. 2024; Ericsson et al. 2022b; Zong et al. 2024) are concepts very similar to CL. Uelwer et al. (2025) define two concepts that are normally considered to be sub-forms of CL as distinct: information maximization (other authors more often use the term feature decorrelation) and teacher-student models (more often referred to as self-distillation) (Gui et al. 2024; Uelwer et al. 2025). The concept of non-contrastive learning is based on joint-embedding that does not require contrastive loss elements and only depends on positive sample pairs, such as BYOL (Abdulrazzaq et al. 2024; Berg et al. 2022; Rani et al. 2023), while most other reviews consider these models as CL without negative samples (Gui et al. 2024). (Knowledge) self-distillation requires no negative examples and is based on maximizing the similarity between the predictions of a complex teacher and a simpler student model (Huang et al. 2022; Khan et al. 2025b). Other authors consider it as a sub-form of CL (e.g. Gui et al. (2024)). The terms auxiliary pretext task (Albelwi 2022; Rani et al. 2023), context-based SSL (Gui et al. 2024; Xu 2021; Zhang et al. 2025b), pixel-to-scalar (Chowdhury et al. 2021), innate relationship (Huang et al. 2023), pseudo-labeling (Liu et al. 2024c) and transformation prediction (Ericsson et al. 2022b) are used as synonyms or similar concepts to predictive SSL. In pixel-to-scalar (Chowdhury et al. 2021), a model learns to map the pixels of the input image to a single value or low-dimensional vector instead of reconstructing the entire image, such as in rotation or jigsaw. Huang et al. (2023) define innate relationship as pretraining with a "hand-crafted task". Similarly, Liu et al. (2024c) establish pseudo-labeling, where distinctive regions are automatically identified to create artificial labels for unlabeled data. In transformation prediction (Ericsson et al.

2022b), the input images with a canonical view are first transformed and then a model is trained to predict this transformation to learn structural and temporal relationships within the data. "Context-based" SSL (Jin et al. 2020) includes tasks based on spatial and temporal context structure as well as context similarity. Jin et al. (2020) define "free semantic label-based methods" with automatically generated labels. According to Jing and Tian (2020), cross modal SSL makes use of different types of input data while Zhang et al. (2023c) define multi-SSL as methods that combine multiple pretext tasks. Graph-based SSL (Zhang et al. 2025b) is optimized for graph data.

Instead of generative learning, some publications exclusively/additionally consider MIM (Chen et al. 2024b; Ciocarlan et al. 2024) (also named masked prediction (Ericsson et al. 2022b; Zong et al. 2024)) as it is the most important type of generative SSL. Hondru et al. (2025) furthermore distinguish between reconstruction- and contrastive-based MIM. Pixel-to-pixel SSL (Chowdhury et al. 2021) trains an autoencoder to reduce the difference between the reconstructed output and the input on a per-pixel basis. Ohri and Kumar (2021) distinguish between three generative approaches: "reconstruction from a corrupted or partial image" including denoising autoencoder and image inpainting, "reconstruction from an altered view of image" including colorization, Split-Brain and BigBiGAN as well as "image generation" based on GAN. Huang et al. (2023) define models like MAE or BEiT as "self-prediction", where parts of the input are either masked or augmented to train the model how to reconstruct missing parts. Consequently, the classification into different paradigms often does not differ significantly, as different terms are used for very similar concepts, or subcategories are understood as separate paradigms.

4.2 Multimodal self-supervised learning

Multimodal SSL learns effective representations from unlabeled data across different modalities (such as image, video, text or audio) and can make use of one modality to provide supervision signal for another (Bai et al. 2025; Goyal 2022; Zong et al. 2024). Consequently, it is one of the most relevant trends regarding SSL. Several authors address multimodal SSL for specific applications, e.g. medicine (Fedorov et al. 2024), recommendation systems (Xu et al. 2025), action recognition (Yang et al. 2023d) or remote sensing (Bai et al. 2025). Multimodal SSL can improve scalability and generalizability of SSL and learn richer representations than unimodal SSL (Deldari et al. 2022; Thapa 2022). As the main advantage of SSL is the ability to train with unlabeled data, the concept of multimodal training introduces new challenges, as multimodal data differs significantly from uni-modal data, on which most SSL paradigms and frameworks are based (Zong et al. 2024). Multimodal approaches can be contrastive (Dufumier et al. 2025; Jia et al. 2021; Mi et al. 2024; Radford et al. 2021; Yang et al. 2023d), include masked modeling methods (Bachmann et al. 2022; Lu et al. 2022; Mizrahi et al. 2023) or are based on clustering, where pseudo-labels are assigned as supervisory signals for other modalities (Dufumier et al. 2025). Multimodal SSL can be realized by pairing multimodal pretraining with a fusion architecture or combining independently pretrained uni-modal models via stitching (Zong et al. 2024). Consequently, approaches can be distinguished by the moment of multimodal integration as immediate, intermediate or late fusion (Deldari et al. 2022). Immediate integration learns representations jointly with unified encoders and is normally more robust regarding missing modalities but inefficient regarding retrieval tasks (Deldari et al. 2022; Zong et al. 2024).

Intermediate fusion merges the outputs of modality-specific encoders into a unified space, which can blur modalities but combines the data on a feature level (Boulahia et al. 2021; Li and Tang 2025). Late fusion only merges at the decision level: it is easy to scale but difficult to model modality interactions fully, resulting in lower performances regarding downstream tasks with complex cross-modal requirements (Li and Tang 2025; Zong et al. 2024). If individually pretrained uni-modal models are used, they can either be frozen or fine-tuned (or anything in between including e.g. partial freezing of chosen layers) (Zhai et al. 2022b). Freezing powerful uni-modal encoders can stabilize multimodal models and reduce training costs (Li et al. 2023c; Zhai et al. 2022b).

A central research challenge is understanding multimodal interactions and how to learn novel, task-relevant information that is not included in any single domain (Liang et al. 2024). Multimodal SSL must cope with perfectly paired, weakly paired, and unpaired data, regimes requiring different learning objectives and architectures (Zong et al. 2024). Treating every modality as an individual view can result in learning shared representations via contrastive objectives which only include redundant cross-modal information (Dufumier et al. 2025). Instead, CoMM aligns "multimodal representations by maximizing the mutual information between augmented versions of these multimodal features" to model interactions "beyond redundancy" (Dufumier et al. 2025). ImageBind demonstrates that image-paired data is sufficient to bind six different modalities together (Girdhar et al. 2023). CLIP applies CL on image-text pairs for representation learning using web-scale and noisy data and has been one of the most influential and prototypical multimodal models (Radford et al. 2021; Wei et al. 2023a). Pre-trained CLIP models have been utilized as teachers to supervise other models due to the strength of their visual representation (Girdhar et al. 2023; Liu et al. 2023f; Peng et al. 2022; Wei et al. 2022b) and as a tokenizer for many MIM models (Dong et al. 2023b; Fang et al. 2023b; Wei et al. 2022c; Yang et al. 2023a; Zhang et al. 2023a). Tokenizer design is a fundamental for MIM (Du et al. 2024), using CLIP extends MIM to multimodal semantics and allows to use "the text space as target space for the reconstruction objective" (Hondru et al. 2025). Multimodal MIM can profit from converting all modalities into discrete tokens instead of using modality-specific nuances (Mizrahi et al. 2023) and mulit-modal masking strategies (Bachmann et al. 2022; Mizrahi et al. 2023). Using pseudo-labels is furthermore one strategy to make frameworks widely applicable and multimodal models seem to increase transferability when being trained with all modalities (Bachmann et al. 2022; Mensink et al. 2022). According to Girdhar et al. (2023), many aspects, such as data augmentation strategy and strength or batch size, are domain-specific in multimodal SSL. Likewise, the choice of pretext objective is crucial and "interacts with fusion and data alignment, shaping what information is captured across modalities" (Zong et al. 2024). Being a rather recent approach, there is a considerable amount of research needed regarding design choices specific to multimodal SSL. For design choices that have been researched, the findings are discussed in the respective chapters (see Sect. 10.1, 10.4 and 11.3).

4.3 Large model pretraining

SSL is particularly attractive for large-scale pretraining because it removes the need for ground-truth labels, enabling training on massive, diverse datasets and high-capacity architectures (Campanella et al. 2025, 2024; Chen et al. 2024b; Xiao 2024). Recently, large model pretraining has e.g. been applied to pathology (Campanella et al. 2024, 2025), diag-

nostics (Tayebi Arasteh et al. 2024), person re-identification (Hu et al. 2025) and image segmentation (Kirillov et al. 2023). Furthermore, the use of ViT is increasingly common in the context of SSL (Khan et al. 2025b). Such transformer architectures perform better with large datasets and models (Goldblum et al. 2023). The combination of SSL and deep networks is essential for the rise of foundation models, that can be adapted to tasks similar to the one they have originally been trained on and are predominantly multimodal (Awais et al. 2025; Bommasani 2023). According to Awais et al. (2025), vision foundation models can be distinguished into textually prompted (e.g. CLIP (Radford et al. 2021), MaskCLIP (Dong et al. 2023b), EVA (Fang et al. 2023b)), visually prompted (e.g. SAM (Kirillov et al. 2023)) and other models (e.g. ImageBind (Girdhar et al. 2023)). Scaling up SSL is especially attractive for the pretext task, as this can improve pretext feature quality and generalization (El-Nouby et al. 2021; Hu et al. 2025; Khan et al. 2025b; Rani et al. 2023). Downstream performance mostly increases when scaling up model capacity and the gains due to larger pretraining datasets seem to increase with model size (Azizi et al. 2021; Singh et al. 2023). Bigger models also perform better given few labeled downstream samples and fine-tuning (Chen et al. 2020c) and are robust regarding distribution shifts (Azizi et al. 2021). At the same time, larger datasets often require more complex pretext tasks (Chen et al. 2024b). Initialization seems to be of considerable importance for large-scale pretraining, so self-supervised pre-pretraining can be useful (Singh et al. 2023). Also, specific data augmentation strategies should be designed for large-scale datasets (Dosovitskiy et al. 2020; Zhang and Ma 2022). Limits of large model pretraining are the enormous computational resources and datasets required, also taking into consideration that computational costs grow faster than performance (Fernandez et al. 2025). Therefore, designing more efficient self-supervised pretraining tasks is important. Also, domain shifts from pretraining to the downstream task remain an issue (see Sect. 8.4), although it is more likely that useful representations are learned from large pretrained networks that have been trained with diverse data (Lavoie et al. 2025). Consequently, large model pretraining does not only require large datasets and large model capacities, but also many other design choices have to be in accordance. The topic is addressed in more detail in Sect. 7 (focus on network size) and Sect. 8.3 (focus on dataset size). Also, in-domain pretraining of SSL in low-data regimes (50k-300k images) can outperform large-scale pretraining (Konstantakos et al. 2025), which is addressed in Sect. 8.4.

5 Self-supervised learning paradigms

While all SSL approaches share the idea of a pretext task that can extract semantically meaningful visual information for the downstream task, there are three major types with distinct learning paradigms: predictive, contrastive and generative SSL (Shurrab and Duwairi 2022; Wu et al. 2023; Wang et al. 2022a; Xie et al. 2023c; Yu et al. 2024). An overview of the three paradigms and their main differences can be found in Fig. 3. The first approaches to SSL were mostly predictive: self-supervision signals were generated from the data itself by predicting mostly geometric transformations such as Relative Position (Rel. Position) (Doersch et al. 2015), Rotation (Gidaris et al. 2018), or Jigsaw puzzle (Noroozi and Favaro 2016). Thereby, these approaches learn the visual structure as well as the contextual semantics and generate pseudo-labels which enable supervised pretraining (Wang

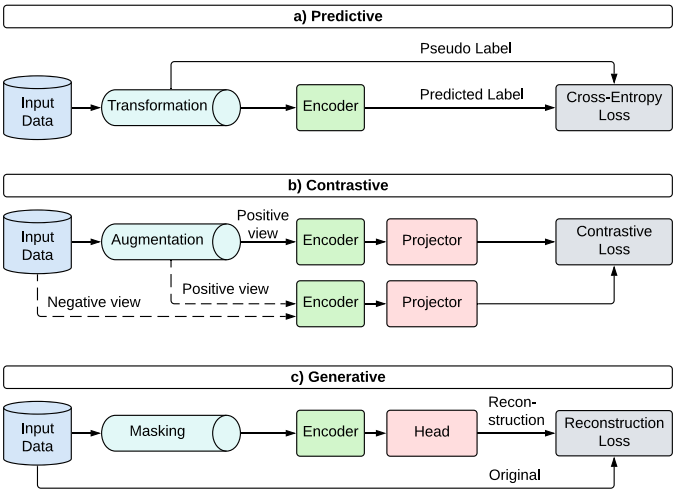

Fig. 3 Schematic comparison of predictive, contrastive and generative SSL based on Liu et al. (2023e) and Wu et al. (2023). Predictive SSL relies on pseudo-labels, contrastive SSL on contrasting views and generative SSL on masking and reconstruction

et al. 2023a). Since 2020, contrastive SSL has made considerable progress with models such as SimCLR, BYOL or MoCo (Jaiswal et al. 2020). Data augmentation plays a key role to generate multiple views which can be contrasted (Huang et al. 2024b). While positive views originate from the same instance, negative views are derived from other instances. The primary objective is to learn similarities and dissimilarities between the data by optimizing the contrastive loss (Chen et al. 2020b). The need for negative samples also results in high computational costs as pairwise comparisons is required, so several methods without negative samples have been developed, often employing architectural innovations (e.g. asymmetry and predictor networks) (Tendle and Hasan 2021; Tian et al. 2020a). Until the combination of ViT with masked image modeling (MIM) in 2021 (Bao et al. 2021), contrastive SSL overshadowed generative approaches due its superior performance (Li et al. 2024a). MIM masks and reconstructs patches of the input data and thereby learns intra-data information by comparing their solution with the original (He et al. 2022; Marks et al. 2025). They focus more on details than contrastive SSL (Liu et al. 2023e; Park et al. 2023; Wu et al. 2023). Recently, hybrid approaches emerged to combine the advantages of two or more paradigms (Wang et al. 2024c).

5.1 Predictive methods

By transforming the pretraining data, predictive models generate pseudo-labels as supervision signals based on the intrinsic data structure (Ericsson et al. 2022b; Yu et al. 2024; Wu et al. 2023). Even comparably simple transformations can result in robust and effective pretext representations (Shurrab and Duwairi 2022). The pseudo-labels are predicted in the pretext task typically using a classification or regressing problem, striving to learn features useful for the downstream task (Kwasigroch et al. 2020; Shurrab and Duwairi 2022; Wu et al. 2023; Wang et al. 2022a). The extracted contextual information depends highly on the pretext task design (Wang et al. 2022a). Typical predictive tasks include (Shurrab and

Duwairi 2022; Zhang et al. 2023c): Relative Position (Rel. Position) (Doersch et al. 2015), Jigsaw (Noroozi and Favaro 2016), Rotation (Feng et al. 2019; Gidaris et al. 2018; Imran et al. 2018), Exemplar (Dosovitskiy et al. 2015), Relative Patch Location (RPL) (Taleb et al. 2000), Rubik's cube+ (RC+) (Zhu et al. 2020) and Superpixel-based pseudo-labels (SPL) (Ouyang et al. 2020). See Table 5 for more examples. Pretext task design is mostly hand-crafted and requires specific domain or expert knowledge about the data structure (Albelwi 2022; Jaiswal et al. 2020; Shurrab and Duwairi 2022). While having been vastly outperformed by contrastive and generative methods (Chen et al. 2020a, b; Grill et al. 2020; Tian et al. 2020a, b; Wang et al. 2023a), predictive SSL is good at focusing on the object of interest (Zhang et al. 2023c) and still applied in medical imaging (Anton et al. 2022; Park and Ryu 2024; Shurrab and Duwairi 2022; Wang et al. 2023a; Zhang et al. 2023c).

5.2 Contrastive methods

Contrastive SSL, also known as contrastive learning (CL), encourages learning representations that can discriminate between different categories by grouping similar samples closer and dissimilar ones further apart (Babaev et al. 2022; Park et al. 2023; Tendle and Hasan 2021). Consequently, inter-class separability and intra-class compactness can be achieved (Wang et al. 2023a). Positive samples are generated by transforming the original data via data augmentation techniques, all other original samples from the batch or whole dataset are considered as negative samples (Falcon and Cho 2020; Huang et al. 2024b; Jaiswal et al. 2020; Ozbulak et al. 2023). Consequently, positive–negative pairs are not created explicitly from the ground-truth target data but implicitly by data augmentation and sampling (Falcon and Cho 2020; Wang and Qi 2023). The use of negative views stabilizes training, but requires large memory resources and complex queuing systems (Chen et al. 2020b; He et al. 2020; Liu et al. 2022a). Approaches without negative views are more efficient, but can result in trivial solutions, meaning that the training data is mapped into a single point (Liu et al. 2022a) or few points (Zhang et al. 2024e) in the feature space, a "shrinkage of the embedding vectors toward zero" (Gui et al. 2024), to the same constant

Table 5 Overview of important predictive models, including the name of the predictive method (Method), whether multiple or one single pretext task is used (Multiple) and a description of the respective pretext tasks (Pretext Task). Exemplar applies the listed transformations on patch level

Method	Multiple	Pretext task
CDJP (Kim et al. 2018)	Multiple	Jigsaw, inpainting, colorization
Colorization (Zhang et al. 2016)	Single	Colorization
Exemplar (Dosovitskiy et al. 2015)	Multiple	Translation, scale, rotation, contrast, color
Inpainting (Pathak et al. 2016)	Single	Inpainting (context encoder)
Jigsaw (Noroozi and Favaro 2016)	Single	Jigsaw puzzle
Models Genesis (Zhou et al. 2019a)	Multiple	Non-linear, local-shuffle, out-/inpainting
MTL (Zhang et al. 2024b)	Multiple	Relative location, orientation regression
SPL (Ouyang et al. 2020)	Single	Superpixel pseudo-labels
RC+ (Zhu et al. 2020)	Multiple	Cube ordering, rotation, masking
Rotation (Gidaris et al. 2018)	Single	Rotation
RPL (Taleb et al. 2000)	Single	Relative 3D patch location
Rel. position (Doersch et al. 2015)	Single	Predicting relative positions
OE (Poursaeed et al. 2020)	Single	3D rotation

vector or a small subspace (Giakoumoglou et al. 2025). Consequently, no useful information can be learned from the data (Giakoumoglou et al. 2025). Contrastive methods can hence suffer from complete representation or dimensional collapse (Giakoumoglou et al. 2025; Jing et al. 2021) and require additional collapse-preventing mechanisms, such as an extra predictor, stop-gradient, additional clustering or decorrelation (Jing et al. 2021; Liu et al. 2022a; Zhang et al. 2024e).

Following the encoding of the views, a projector maps the input space into the embedding space where contrastive loss is applied in order to measure the closeness of two embeddings (Assran et al. 2023; Balestriero et al. 2023). The contrastive loss is a key component of CL and enables "the model to learn the difference between similar instances and the similarity between dissimilar instances" (Hu et al. 2024b; Wang and Liu 2021). Noise Contrastive Estimation (NCE) loss (Gutmann and Hyvärinen 2010) and InfoNCE loss (Oord et al. 2018) are the most common contrastive loss options (Hu et al. 2024b). NCE loss is similar to softmax but distinguishes between true data and negative samples to avoid conservative predictions given a large number of categories (Hu et al. 2024b). NCE assumes that negatives (which are also referred to as noise) are independent from the positive anchor sample (Uelwer et al. 2025). InfoNCE is the most common contrastive loss and an optimization of NCE which maximizes Mutual Information (MI) between positive pairs and minimizes MI between negative pairs (Hu et al. 2024b). Several variations of InfoNCE furthermore have been developed over time, e.g. NT-Xent by SimCLR which includes temperature scaling (Chen et al. 2020b; Khan et al. 2025b). Other contrastive losses are ProtoNCE (Li et al. 2021e), Nearest Neighbour Contrastive Learning loss (NNCLR) (Dwibedi et al. 2021), Contrastive AUC (Sharma et al. 2023), Adversarial Contrastive Loss (AdCo) (Hu et al. 2021), Augmentation Robust loss (AR) (Zhao et al. 2023) or alignment loss (Align) (Lavoie et al. 2025). The most commonly used non-contrastive loss functions are mean squared error (MSE), which is both influenced by positive and negative samples, and cross-entropy loss (CE), which can measure the difference between the predicted and the real distribution and is very common for classification problems (Lavoie et al. 2025). Alternatives are logistic loss, negative cosine similarity (NCS), Barlow Twin loss (BTL) (Zbontar et al. 2021), MixUp regularization loss (MixUp) (Bandara et al. 2023), Variance-invariance-covariance loss (VIC) (Bardes et al. 2021), Location-based and feature-based loss (LF) (Bardes et al. 2022), Centroid loss (Centroid) (Estepa et al. 2023), Redundancy Reduction loss (RR) derived from Barlow Twins (BT) (Bandara et al. 2023) and Faster R-CNN loss (Benigmim et al. 2024).

CL is very successful in CV with discriminative downstream tasks (Lee et al. 2022a; Liu et al. 2023e; Ozbulak et al. 2023; Tendle and Hasan 2021) due to the alignment of features from positive pairs, the distribution uniformity on the feature-hyper-sphere (Wang and Isola 2022), the divergence of class centers and the concentration of augmented data (Huang et al. 2021). Different subtypes vary in loss function, architecture, the generation of positive as well as negative views (Wang et al. 2023a) and can be classified into approaches with negative samples, cluster discrimination, self-distillation and feature decorrelation (Balestriero et al. 2023; Jaiswal et al. 2020; Ozbulak et al. 2023; Wang et al. 2023a, 2022a). Figure 4 illustrates those subtypes and the following sections explain them in detail. Table 6 presents an overview of important CL models.

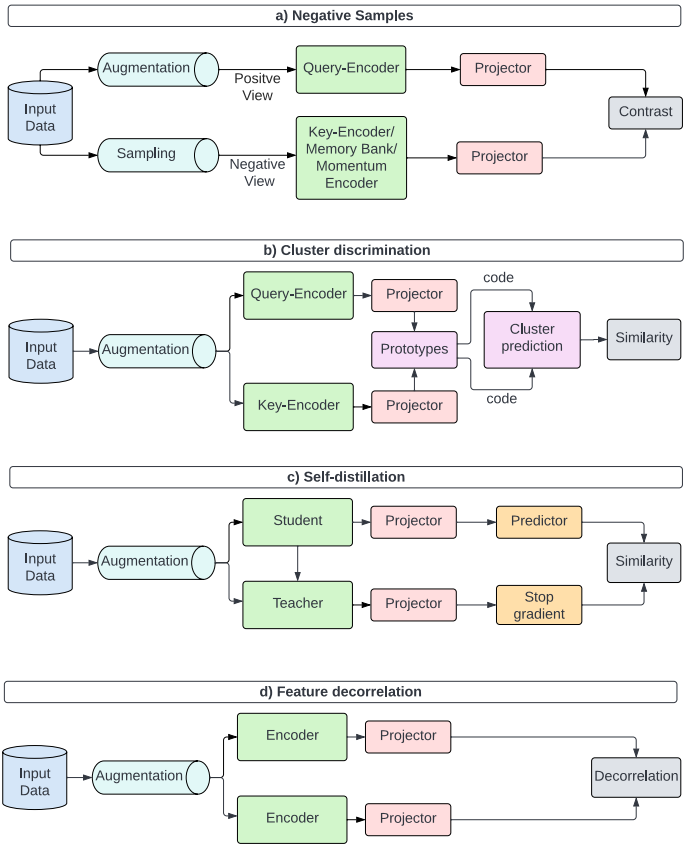

Fig. 4 Schematic comparison of the different types of contrastive SSL inspired by Wang et al. (2022a) and Jaiswal et al. (2020). Methods with negative samples can be distinguished by the different types of encoders they employ. Cluster discrimination relies on cluster prediction, some including prototypes. Self-distillation and feature decorrelation models typically use an asymmetric teacher-student structure. Self-distillation approaches can furthermore include predictors and stop gradient

5.2.1 Contrastive methods with negative samples

CL with negative samples is based on instance-discrimination as pretext task, where each data sample is considered as an independent instance and distinct views of one sample are generated via data augmentation (Gui et al. 2024; Hu et al. 2024b). By distinguishing instances without relying on semantic categories, the model learns similarities among instances instead of classes (Hu et al. 2024b). As every instance is treated as a discrete class, the model is very generalizable (Chen et al. 2020c; Jaiswal et al. 2020; Tendle and Hasan 2021). Including negative samples facilitates to discern the feature similarities within the positive views by systematically distancing the positive pairs from the negative samples within the embedding space (Chen et al. 2020b; He et al. 2020). Harder samples are generally beneficial, as they increase the difficulty of the pretext task (Dong et al. 2024). CL with negative samples can be divided according to the respective strategy for collecting negative samples (Jaiswal et al. 2020). End-to-end approaches, such as SimCLR, individually end-

Table 6 Overview of CL models including type, properties and loss function. Clustering (C) is either offline or online, while Self-Classifier uses a single-stage approach. Prototypes group similar instances or features to avoid pairwise comparison. Negative samples (N) are managed with memory banks, Nearest Neighbor (kNN), batches or momentum encoders or include adversarial samples (Adv). Self-distillation (SD) relies on Student/Teacher or Siamese architectures. Feature decorrelation (FD) is based on redundancy reduction or variance-invariance-covariance regularization

Method	Type	Properties	Loss
CoKe (Qian et al. 2022)	C	Online	Softmax, CE
DC (Caron et al. 2018)	C	Offline	Logistic
DC-v2 (Caron et al. 2021a)	C	Offline	Logistic
PCL/PCL-v2 (Li et al. 2021e)	C	Offline, Prototype	InfoNCE, ProtoNCE
Self-Classifier (Amrani et al. 2022)	C	Single-Stage	CE
SwAV (Caron et al. 2021a)	C	Online, Prototype	CE
GC (Wang et al. 2024a)	C, SD	Student/Teacher	InfoNCE, CE
BT (Zbontar et al. 2021)	FD	Redundancy	BTL
Mixed BT (Bandara et al. 2023)	FD	Redundancy	BTL, MixUp
VICReg (Bardes et al. 2021)	FD	Variance	VIC
VicRegL (Bardes et al. 2022)	FD	Variance	VIC, LF
All4One (Estepa et al. 2023)	FD	Redundancy	NNCLR, Centroid, RR
AUC-CL (Sharma et al. 2023)	N	Batch	Contrastive AUC
SimCL (Yang et al. 2022)	N	Batch	InfoNCE
SimCLR (Chen et al. 2020b)	N	Batch	InfoNCE
SimCo (Zhang et al. 2022a)	N	Batch	Hardness-aware InfoNCE
DnC (Tian et al. 2021)	N	Batch	InfoNCE
MoCLR (Tian et al. 2021)	N, SD	Batch	InfoNCE
AdCo (Hu et al. 2021)	N	Memory bank, Adv	AdCo
CaCo (Wang et al. 2023e)	N	Memory bank, Adv	InfoNCE
NNCLR (Dwibedi et al. 2021)	N	Memory bank, kNN	NNCLR
InstDisc (Wu et al. 2018)	N	Memory bank	NCE
MoBY (Xie et al. 2021a)	N	Memory bank	InfoNCE
PIRL (Misra and Maaten 2020)	N	Memory bank	InfoNCE
SimCLR-v2 (Chen et al. 2020c)	N	Memory bank	InfoNCE
ArCL-MoCo (Zhao et al. 2023)	N	Memory bank	AR
Mugs (Zhou et al. 2022c)	N	Memory buffer	InfoNCE, CE
MoCo (He et al. 2020)	N	Momentum encoder	InfoNCE
MoCo-v2 (Chen et al. 2020a)	N	Momentum encoder	InfoNCE
MoCo-v3 (Chen et al. 2021c)	N	Momentum encoder	InfoNCE
BYOL (Grill et al. 2020)	SD	Student/Teacher	MSE
DINO (Caron et al. 2021b)	SD	Student/Teacher	Softmax, CE
DINO-MC (Wanyan et al. 2024)	SD	Student/Teacher	Softmax, CE
DINO-v2 (Oquab et al. 2024)	SD	Student/Teacher	Softmax, CE
DT (Lavoie et al. 2025)	SD	Student/Teacher	Faster R-CNN, Align
SimSiam (Chen and He 2021)	SD	Siamese	NCS

to-end update a query encoder (trained on the original) and a key encoder (trained on the positive and negative views), aiming to make positive samples more similar and negative ones more dissimilar from the original (Chen et al. 2020b; Jaiswal et al. 2020). For these models, the batch size is of great importance, as it controls the amount of available negative samples and mitigates optimization errors (Chen et al. 2021c, 2022e; Hu et al. 2024b).

Instead of scaling up the batch size, memory banks can act as separate dictionaries keeping a large number of feature representations of negative samples (Jaiswal et al. 2020), like PIRL (Misra and Maaten 2020). The computational effort of maintaining the memory bank is a considerable disadvantage (Chen et al. 2020b). Momentum encoders include a queue of encoded negative samples consisting of the current mini-batch and the oldest dequeued mini-batch (He et al. 2020; Jaiswal et al. 2020; Liu et al. 2023e). The most prominent example is MoCo (He et al. 2020). While instance-discrimination-based approaches with negative samples prevent network collapse by using negative samples, they can suffer from representation dimension collapse (Jing et al. 2021; Hua et al. 2021). Models with negative samples come with high computational costs as pairwise comparison is required (Tendle and Hasan 2021).

5.2.2 Contrastive methods with cluster discrimination

Clustering-based approaches learn by distinguishing between samples having been assigned to different clusters in the embedding space (Tendle and Hasan 2021). First, the encoded representations are clustered, then the clustering is predicted (Zong et al. 2024). These assignments can act as supervision signals (Zong et al. 2024). Cluster centers can be used as proxies for both negative and positive samples, which solves the issue of using negative samples which are semantically too similar to positives (Hu et al. 2024b; Li et al. 2021e). Cluster discrimination in general maximizes "the entropy of the average embedding" (Assran et al. 2023). DeepCluster (Caron et al. 2018) e.g. makes use of k-means clustering to generate discriminable pseudo-labels, SwaV (Caron et al. 2021b) learns features by swapping assignments between positive views and PCL (Li et al. 2021e) employs prototypes (Gui et al. 2024). Some models have to deal with general clustering challenges, e.g. determining the right cluster number, the domination of large clusters, offline training, trivial solutions or simultaneous feature learning (Ozbulak et al. 2023; Wang et al. 2023a).

5.2.3 Contrastive methods with self-distillation

Self-distillation models are designed to work without negative samples by maximizing the similarities between two positive views (Gui et al. 2024). Important self-distillation models are BYOL (Grill et al. 2020), SimSiam (Chen and He 2021), FastSiam (Pototzky et al. 2022), DINO (Caron et al. 2021b), and DINO-v2 (Oquab et al. 2024). The architecture includes a teacher and a student network with either an asymmetric learning role or architecture (e.g. asymmetric siamese architectures) to allow KT between the networks without needing negative samples (Wang et al. 2022a). The pretrained, frozen teacher network is both larger and more powerful and transfers its knowledge to the smaller student network (Giakoumoglou et al. 2025). The student is trained by predicting the representation of the teacher (Grill et al. 2020; Liu et al. 2022a; Ozbulak et al. 2023). The views are inserted into the different encoders and mapped to each other using a predictor (Balestriero et al. 2023). Self-distillation methods can be distinguished by their strategy to prevent model collapse (Balestriero et al. 2023). An additional predictor after the student and stop gradient operations of the teacher help, but the latter is significantly more important (Chen and He 2021; Richemond et al. 2023). Stop gradient is integrated into the teacher network and enables asymmetric learning by restricting gradient flow to the student network, so that the teacher is not affect by the student's

predictions (Jang et al. 2023; Ong et al. 2025). This separation improves training stability as interference between the two sub-networks is significantly reduced, thereby avoiding collapses due to trivial solutions (Giakoumoglou et al. 2025). Some self-distillation approaches (e.g. BYOL, SimSiam, FastSiam) also integrate a predictor on the student branch to ensure asymmetry, which helps the student network to lean more information and makes collapsed solutions less likely (Grill et al. 2020). BYOL's linear predictor acts as an "an approximate orthogonal projection as well as a spectral expansion operator" (Richemond et al. 2023). It fosters a "decorrelation of linear encoder features" (Richemond et al. 2023) and "implicit variance regularization" (Srinath Halvagal et al. 2023). Predictors can augment the covariance of the features and prevent collapse (Richemond et al. 2023).

5.2.4 Contrastive methods with feature decorrelation

Maximizing the information of decorrelated embeddings is a successful approach neither requiring negative samples nor asymmetric architectures (Ozbulak et al. 2023; Zbontar et al. 2021). The main idea is that the output feature dimensions of the model are uncorrelated (Maitra et al. 2025). According to information theory, learning decorrelated features fosters good representations (Gui et al. 2024; Zbontar et al. 2021). The methods aims to identify the relationship between two variables by studying their cross-covariance matrices (Balestriero et al. 2023). After the encoder, the so called expander projects the representations into the embedding space (Bardes et al. 2021). Bardes et al. (2021) use the term "expander" instead of projector due to the increase of representation dimensionality. Its purpose is removing information that distinguish the representations of the two branches and non-linearly increasing the representation dimensionality to ensure that decorrelating the embedding variables can reduce the dependencies (Bardes et al. 2021). Collapse can be prevented by innovative loss functions as feature deccorlation reduces dimensional collapse (Hua et al. 2021; Ozbulak et al. 2023). Consequently, the models can be distinguished according to these loss functions. Important examples for feature decorrelation are BT (Zbontar et al. 2021), VicReg (Bardes et al. 2021) or VicRegL (Bardes et al. 2022).

5.3 Generative methods

Generative approaches are based on intra-data information (Liu et al. 2021b) and learn appropriate distributions according to the data in the pixel space (Ozbulak et al. 2023). The key idea is that a representation which enables a high-quality reconstruction of the input distribution includes the required categorical information to discriminate between samples (Tendle and Hasan 2021). Generative models are trained to learn latent features (Shurrab and Duwairi 2022) and can "recover the original data distribution without assumptions for downstream tasks", which makes them useful in a wide range of downstream applications (Liu et al. 2023e) and mostly data-agnostic (Kong and Zhang 2023). Still, a highly detailed reconstruction is not necessarily beneficial regarding downstream performance (Newell and Deng 2020). Generative SSL can be computationally expensive, sensitive to outliers and tends to focus on low-level, locally biased features (Li et al. 2024a; Liu et al. 2023e; Ozbulak et al. 2023; Xie et al. 2023a; Zhang et al. 2023e). The most important generative approaches are Generative Adversarial Networks (GAN), diffusion models,

MIM or combinations. MIM is dominant for discriminative downstream tasks. An overview of important models is presented in Table 7, a detailed categorization in Fig. 5.

5.3.1 Generative adversarial networks

GANs are based on adversarial learning, which can be understood as a combination of generative and contrastive learning and is a good choice for implicit distributions (Liu et al. 2021b). GANs can either generate images using the complete input or recover with partial input (Liu et al. 2023e). The latter option is preceded by a predictive pretext tasks (such as rotation (Chen et al. 2019b), jigsaw (Baykal and Unal 2020; Baykal et al. 2022)), colorization (Treneska et al. 2022) or inpainting (Zhou et al. 2023d), rendering these models more similar to MIM (Liu et al. 2023e). While the encoder projects the input into a latent space, the decoder re-projects the latent features into the original space, striving to reconstruct the original data (Goodfellow et al. 2020; Schiappa et al. 2022). Finally, a discriminator is trained to distinguish the generated version from the original (Goodfellow et al. 2020; Schiappa et al. 2022). The application of ViT backbones (Yu et al. 2022; Zhang et al. 2024c; Zhou et al. 2023b), self-labeling (Lučić et al. 2019), label-augmentation (Hou et al. 2021) and better downstream transferability (Donahue et al. 2016) have further improved self-supervised GANs (Ozbulak et al. 2023). Still, problems such as mode collapse, low performance in discrete settings, a lack of backbone architecture variability and scalability issues have so far prevented a breakthrough of GANs for SSL (Liu et al. 2023e; Ozbulak et al. 2023).

5.3.2 Diffusion models

Diffusion models are highly successful likelihood-based generative models that outperform GANs on image synthesis (Ayromlou et al. 2025; Dhariwal and Nichol 2021; Li et al. 2023a). These models apply the re-weighted variational lower-bound as their training target, which enables them to learn how to transform an image into noise and vice versa (Ayromlou et al. 2025; Ho et al. 2020). Due to their long training times, they are usually not used for discriminative tasks (Yamaguchi and Fukuda 2023), which is why they are excluded from our considerations.

5.3.3 Masked image modeling

Recent generative methods for discriminative tasks are mostly based on MIM (Hondru et al. 2025; Ozbulak et al. 2023). The modeling goal is to deduce a latent variable that represents the common information between visible and invisible patches of an image, using only the visible ones (Kong et al. 2023). Reconstruction should not increase complexity to a degree which is unnecessary for categorization (Tendle and Hasan 2021), so MIM is most suitable for applications where detailed knowledge (Krishnan et al. 2022) and low-level abstraction is required (Liu et al. 2023e). BEiT was the first self-supervised MIM approach that showed a better downstream performance than state-of-the-art CL methods (Zhang et al. 2023e). In general, the performance of MIM has approached or even exceeded that of CL (Bao et al. 2021; Chen et al. 2020d, 2024a; Hou et al. 2022; Li et al. 2024a; Oquab et al. 2024; Woo et al. 2023). Pretrained tokenizers (Zhou et al. 2022b) and knowledge distillation from pre-

Table 7 Overview of generative SSL methods by type: MIM, combinations of MIM and GAN (M+G), GAN, diffusion models (Diff) or combinations of diffusion model and MIM (D+M). Most approaches are based on autoregression (AR) or autoencoders (AE). MIM uses features (F), Tokenizer (T), Pixels (P) or a Combination (C) as reconstruction target (RT). DeepMIM uses a hybrid (H) approach which is applicable to various RT. The most common masking strategy (Mask) is random masking

Method	Type	AR/AE	RT	Mask
A2MIM (Li et al. 2023d)	MIM	AE	F	Random
AIM (El-Nouby et al. 2024)	MIM	AR	T	Raster
AutoMAE (Chen et al. 2023b)	MIM	AE	P	Advers
BeiT (Bao et al. 2021)	MIM	AE	T	Random
BeiT-v2 (Peng et al. 2022)	MIM	AE	T	Random
BeiT-v3 (Wang et al. 2022b)	MIM	AE	T	Random
BIM (Luo et al. 2023)	MIM	AE	F	Block
BootMAE (Dong et al. 2022)	MIM	AE	C	Block
CAE (Chen et al. 2024a)	MIM	AE	F	Random
CAE-v2 (Zhang et al. 2023a)	MIM	AE	F	Random
CIM (Fang et al. 2022)	MIM	AE	T	Random
DAGMaN (Jiang et al. 2025)	MIM	AE	F	Attention
data2vec (Baevski et al. 2022)	MIM	AE	F	Random
data2vec-v2 (Baevski et al. 2023)	MIM	AE	F	Multi
DeepMIM (Ren et al. 2025)	MIM	AE	H	Random
DILEMMA (Sameni et al. 2023)	MIM	AE	F	Random
DMJD (Ma et al. 2024)	MIM	AE	F	Disjoint
DropPos (Wang et al. 2023c)	MIM	AE	F	Random
EMAE (Li et al. 2024b)	MIM	AE	P	Parallel
EVA (Fang et al. 2023b)	MIM	AE	F	Block
FastMIM (Guo et al. 2022)	MIM	AE	F	Random
Ge2AE (Liu et al. 2023b)	MIM	AR	F	Random
HPM (Wang et al. 2023f)	MIM	AE	F	Hard
Img2vec (Pan et al. 2023b)	MIM	AE	F	Random
iBot (Zhou et al. 2022b)	MIM	AE	T	Random
iGPT (Chen et al. 2020d)	MIM	AR	P	AR
MAE (He et al. 2022)	MIM	AE	P	Random
MaPeT (Baraldi et al. 2023)	MIM	AR/AE	T	Random
MaskAlign (Xue et al. 2023)	MIM	AE	–	Attention
MaskFeat (Wei et al. 2022a)	MIM	AE	F	Random
MaskGIT (Chang et al. 2022)	MIM	AR	T	Random
mc-BeiT (Li et al. 2022b)	MIM	AE	T	Random
MILAN (Hou et al. 2022)	MIM	AE	F	Attention
MIRAM (Vo et al. 2025)	MIM	AE	F	Multiscale
M-MAE (Tan et al. 2024)	MIM	AE	P	Random
MR-MAE (Gao et al. 2024)	MIM	AE	C	Random
MSG-MAE (Tukra et al. 2023)	MIM	AE	F	Random
MSN (Assran et al. 2022b)	MIM	AE	F	Random
MVMAE (Ji et al. 2023)	MIM	AE	F	Random
PCAE (Li et al. 2022e)	MIM	AE	T	Random
PeCo (Dong et al. 2023a)	MIM	AE	T	Random
PixMIM (Liu et al. 2023d)	MIM	AE	F	Simple
RandSAC (Hua et al. 2022)	MIM	AR	T	Random
RC-MAE (Lee et al. 2022b)	MIM	AE	P	Random

Table 7 (continued)

Method	Type	AR/AE	RT	Mask
SdAE (Chen et al. 2022c)	MIM	AE	F	Random
SiameseIM (Tao et al. 2023)	MIM	AE	F	Random
SimMIM (Xie et al. 2022b)	MIM	AE	P	Random
SMIT (Jiang et al. 2022b)	MIM	AE	C	Random
SparK (Tian et al. 2023)	MIM	AE	P	Random
SplitMask (El-Nouby et al. 2021)	MIM	AE	T	Random
TinyMIM (Ren et al. 2023)	MIM	AE	F	Random
UMMAE (Li et al. 2022d)	MIM	AE	P	Uniform
LVM (Bai et al. 2024)	M+G	AR	T	AR
MAGE (Li et al. 2023a)	M+G	AR/AE	T	Random
SeiT++ (Lee et al. 2024)	M+G	AE	T	Token
GAN-MAE (Fei et al. 2023)	G+M	AE	P	Random
BiGAN (Donahue et al. 2016)	GAN	–	–	–
BigBiGAN (Donahue and Simonyan 2019)	GAN	–	–	–
SRGAN (Ledig et al. 2017)	GAN	–	–	–
DALLE-3 (Betker et al. 2023)	Diff	–	–	–
StableDiffusion (Rombach et al. 2022)	Diff	–	–	–
DiffMAE (Wei et al. 2023b)	D+M	AE	C	Random

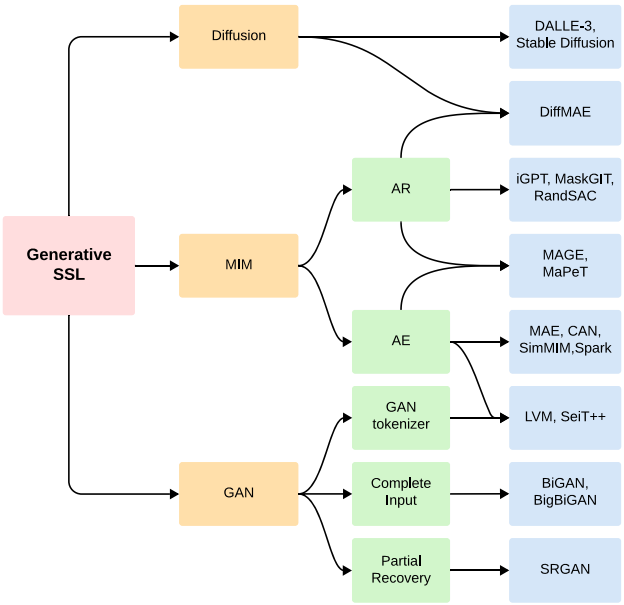

Fig. 5 Overview of different types of generative SSL including diffusion models, MIM and GAN. MIM can be either based on AR or AE methods, while GAN approaches can be characterized by their use of a GAN tokenizer, complete input or partial recovery. Mixed approaches exist. Specific examples of models are highlighted in blue

trained checkpoints (Liu et al. 2023f) helped to bridge the gap between MIM and contrastive LP accuracy (Yao et al. 2022). The success of MIM is directly linked to the development of ViT due to their architectural compatibility regarding masking and reconstruction (Xie et al. 2023a). ViT segments images into patches and thus aligns well with the requirements of MIM regarding masking and predicting parts of the input, enabling a seamless integration of the two approaches (Ozbulak et al. 2023; Zhou et al. 2022b). In contrast to language, images have no inherent tokenizer structure, and pixel-based reconstruction can lead to redundancy and dimensional problems (Li et al. 2024a). In combination with ViT, MIM can train on tokens created from images patches (Ozbulak et al. 2023; Wu et al. 2020). MIM tends to focus on low-level, locally biased features, while ViT use large receptive fields and prioritize global information, consequently the combination boosted generative SSL (Li et al. 2024a; Xie et al. 2023a; Zhang et al. 2023e). Latent MIM with masked reconstruction in latent space could be one possibility to combine the locality of MIM with high-level targets (Wei et al. 2024a). In contrast, CL rather emphasizes high-level features with semantic content and consequently works well with CNNs, which are characterized by a local bias (Li et al. 2024a; Yao et al. 2022). Combining MIM and CNN results in sub-optimal coupling of local bias (Li et al. 2024a).

Most MIM approaches are either based on auto-regressive (AR) or auto-encoder (AE) models (Li et al. 2024a). AR models contain a sequence of regressions, where each prediction depends on previous inputs or outputs (Li et al. 2024a). AE approaches dominate MIM due to their lower self-attention complexity combined with high performance (Li et al. 2024b). They learn by reconstructing the corrupted input, either using denoising AE or masked AE (He et al. 2022). Most MIM models furthermore apply a masking strategy and a reconstruction target (RT) (Li et al. 2024a).

5.4 Hybrid methods

MIM and CL have many opposing but potentially complementary properties. While CL is shape-oriented, MIM is more texture-oriented (Park et al. 2023). Contrastive approaches are very good at focusing on important objects (Lehner et al. 2024), but require vast amounts of training data and are prone to overfitting (Gui et al. 2024). In contrast, pixel-level reconstruction in MIM can avoid patch prediction overfitting (Li et al. 2021a). MIM is more suitable for early layers, while CL can focus on deeper ones (Park et al. 2023). MIM achieves a high degree of detail knowledge (Krishnan et al. 2022; Liu et al. 2023e; Wei et al. 2024a), in turn CL is better at learning high-level semantics (Li et al. 2024a; Yao et al. 2022). CL is better at capturing longer-range global patterns in contrast to MIM (Park et al. 2023). Contrastive objectives foster linear separability of learned representations, but can lead to self-attention collapsing into homogeneity for ViT backbones, which in turn decreases scalability and dense prediction performance (Park et al. 2023). Also, the focus on global patterns in CL is not necessarily beneficial for object recognition and distinguishing images, as it can cause difficulties preserving local information (Park et al. 2023). In turn, MIM can help to preserve the texture and diversity of the learned representations (Park et al. 2023). The addition of contrastive loss improves MIM-representation-learning and linear separability (Huang et al. 2024b; Li et al. 2023a; Zhang et al. 2023b). Consequently, several hybrid forms have been developed to combine the individual advantages. They show promising results regarding representation quality, layer-specific optimization, high performance as

well as two-stage training. Figure 6 provides an overview of hybrid SSL. Table 8 introduces important hybrid models.

Few authors combine MIM only with contrastive InfoNCE loss (e.g. Unified-IO) or contrastive heads (e.g. SplitMask), but many with both elements (e.g. MACRL, CMAE, MaskCLIP, MimCo, SDMAE, ConMIM) (Li et al. 2024a). Furthermore, models like BEiT-v2, BEiT-v3, or CAE-v2 include CLIP, a tokenizer that has been pretrained via CL in a multimodal manner. CMAE, CAN, MimCo or MARCL add a contrastive objective (Yao et al. 2022). MAE-Refiner uses intermediate MIM representation for multiple contrastive heads (Alkin et al. 2025). Instance-based CL profits from incorporating masking as data augmentation (Yang et al. 2023c) (e.g. SIM (Tao et al. 2023)). Standard MIM models also include weak forms of data augmentation, while e.g. adding color jitter degrades MAE performance, indicating that MIM requires different data augmentation strategies compared to CL (Chen et al. 2023a). Due to the diverse possibilities to add contrastive elements to MIM-frameworks, the question arises as to what constitutes a hybrid model. Due to the fact that only adding minor contrastive elements does not change the MIM-framework substantially, this paper only considers approaches as hybrid that include contrastive loss, contrastive head, contrastive objective, data augmentation, masking, reconstruction target, and auto-encoder. These hybrid approaches can combine CL and MIM in a sequential, parallel, or merged manner.

Sequential approaches can make use of the benefits of MIM in earlier layers and CL in deeper layers or they can combine MIM-pretraining with CL. First applying CL and then MIM is possible, but not as common. Parallel approaches combine two distinct MIM and CL models by a weighted loss or shared weights. Merged approaches mutually integrate contrastive and generative elements.

Several publications apply reconstruction in earlier layers and then contrast in deeper ones, as MIM is more effective in earlier layers, where high-frequency information can be extracted, while CL is better for later layers with low-frequency information (Chen et al. 2023c; Jiang et al. 2023; Park et al. 2023). Other sequential approaches first use MIM-

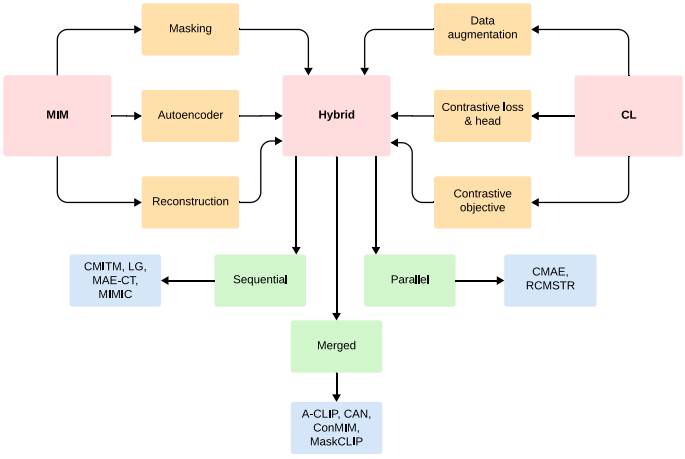

Fig. 6 Overview of hybrid SSL including components derived from MIM (masking, AE, reconstruction) and CL (data augmentation, contrastive loss, head and objective). Hybrid SSL can be categorized into sequential, parallel, merged. Each is illustrated with examples

Table 8 Hybrid methods include masking and RT derived from MIM. Possible RTs include feature (F), pixel (P), tokenizer (T) and combinations (C). Applied losses are InfoNCE, Mixed-supervised contrastive loss (MSC), Kullback–Leibler Divergence (KL), MSE, Nearest Neighbor Alignmen (NNA) (Alkin et al. 2025), CE, or Masked prompting tuning (MPT) (Wu et al. 2025). Heads can be contrastive (C) or contrastive and generative (C+G). Some models use the same augmentation strategies of a popular CL method (SimCLR, BYOL, SimMIM or MoCo). The hybrid type (HT) is merged (M), parallel (P) or sequential (S). The most common contrastive objectives are types of discrimination (discrim.) and negative samples (neg.). SeqCLR is an augmentation strategy developed by Aberdam et al. (2021), while Zhang et al. (2025a) modify random masking

Method	Masking	RT	Loss	Head	CL Objective	Augmentation	HT
A-CLIP (Yang et al. 2023c)	Attention	F	InfoNCE	C	Multimodal, neg	Resize, Crop	M
CAN (Mishra et al. 2022)	Random	P	MSE, InfoNCE	G+C	Joint loss, neg	SimCLR	M
CMAE (Huang et al. 2024b)	Local	P	MSE, InfoNCE	C	View-match, neg	Pixel shift, Color	P
CMITM (Chen et al. 2023c)	Random	P	MSE, InfoNCE	C	Multimodal contrast	Resize, Crop	S
ConMIM (Yi et al. 2023)	Random	P	InfoNCE	C	Patch-discrim., neg	BYOL	M
CoRev2 (Wu et al. 2025)	Random	P	MSE, InfoNCE, MPT	G+C	Instance-discrim., neg	Varies with data	S
LG (Jiang et al. 2023)	Random	P	MSE, InfoNCE	C	Contrastive deep layers	Crop, Flip	S
MACRL (Yao et al. 2022)	Random	P	MSE, InfoNCE	C	Negatives	BYOL	M
MAE-CT (Lehner et al. 2024)	Random	P	MSE, InfoNCE	C	Instance-discrimination	BYOL	S
MaskCLIP (Dong et al. 2023b)	Random	T	MSE, InfoNCE	G+C	Self-dist., multimodal	Central crop	M
MiM (Zhuang et al. 2025)	Multilevel	F	MSE, InfoNCE	G+C	Multi-Level, neg	Resize, Crop	M
MimCo (Zhou et al. 2023c)	Random	F	InfoNCE	C	Contrastive teacher	Crop, Flip, Color	M
MIMIC (Zhang et al. 2024a)	Random	F	MSE, MSC, CE	G+C	Contrastive FT, neg	SimMIM	S
MIM-Refiner (Alkin et al. 2025)	Random	F	NNA	C	Instance-discrim., neg	Varies with data	M
MOMA (Yao et al. 2023)	Random	F	InfoNCE	C	MoCo teacher, neg	MoCo	S/P
RCMSTR (Lin et al. 2024)	Patchblock	P	MSE, InfoNCE, KL	G+C	Relational CL, neg	SimMIM, SeqCLR	P
SDMAE (Mao et al. 2023)	Contextual	C	MSE, InfoNCE	C	Class-token-discrim.	MoCo-v3, MAE	S
SiT (Atito et al. 2022)	Group	P	MSE, InfoNCE	C	Instance-discrim., neg	SimCLR w/o RCD	M
USeLF (Zhang et al. 2025a)	Random	P	MSE, InfoNCE	G+C	Instance-discrim., neg	HSI, Crop, Color	M

pretraining, followed by CL, just like MAE-CT, who apply instance-discrimination via kNN after reconstruction-based pretraining, followed by contrastive tuning, which freezes "the bottom half of the ViT-encoder" and applies "layer-wise learning rate decay" for the remaining part (Lehner et al. 2024). LG combines MIM and contrastive instance-discrimination sequentially (Jiang et al. 2023; Lehner et al. 2024), CMITM applies a cascaded training strategy where the network is first trained with reconstruction loss and then with contrastive loss (Chen et al. 2023c). MIMIC first pretrains a ViT via masked image reconstruction and then fine-tunes via mix-CL (Zhang et al. 2024a). After an image prediction task similar to MAE, the feature encoder and projector of SDMAE use the "class token output by the decoder" for CL (Mao et al. 2023). MOMA can either use MoCo-pretraining distilled to pretrain MAE or vice versa (sequential) or distillation from both to a randomly initialized student model (parallel).

In parallel approaches, the two learning paradigms remain distinct for the majority of the model, typically only being integrated through the application of a loss function. In CMAE, the final learning target is a weighted combination of reconstruction and contrastive loss (Huang et al. 2024b). Alternatively, RCMSTR applies MIM and relational CL in a parallel manner, while the ViT or CNN encoders of the MIM branch and the positive views share weights (Lin et al. 2024).

Merged approaches are more mutually integrated. After masking both views but only augmenting one, CAN applies joint loss training, the masked patches are reconstructed and the pooled patches are contrasted (Mishra et al. 2022). A-CLIP contrasts the output of an image and a text encoder but also allows adding auxiliary SSL tasks (Yang et al. 2023c), while MARCL uses the same encoder output for both objectives (Yao et al. 2022). MaskClip applies feature- and token-wise distillation loss as well as contrastive loss, the contrasted views are multimodal (Dong et al. 2023b). Mimco first applies CL for pretraining and then uses the pretrained encoder as a contrastive teacher in an MIM framework (Zhou et al. 2023c). ConMIM combines a patch-level dynamic dictionary and contrastive denoising, they compare the resulting dictionaries using contrastive loss (Yi et al. 2023). MACRL applies a CL objective to an asymmetric siamese network, wherein one branch has both a higher masking ratio and stronger data augmentations than the other (Yao et al. 2022). MimCo uses a pretrained contrastive model as teacher model in combination with patch-level and image-level reconstruction (Zhou et al. 2023c). Merged hybrid forms are the most diverse of the three subtypes, probably due to the greatest number of design degrees of freedom.

6 Pretext task design

Self-supervised pretraining can extract information relevant for the downstream task, but the results are highly dependent on the specific pretext-downstream-task combination (Jaiswal et al. 2020; Tsai et al. 2021). Ideally, the model should become invariant to the transformations applied in the pretext task while remaining discriminative to characteristics required for the downstream task (Jaiswal et al. 2020; Xiao et al. 2021). The pretext transformations should not alter the training distribution significantly, as this could cause a negative domain shift, especially with a supervised downstream task (Moon et al. 2022; Tian et al. 2022). Furthermore, the applied transformation modes have to be distinguishable to avoid

learning noisy features (Pal et al. 2019). Pretext design should be in accordance with the underlying assumptions of the different pretext task types: instance-based CL requires each sample being a unique semantic example, predictive learning requires all samples having a canonical view, clustering-based CL requires that the data has meaningful similarities, and generative approaches require that parts of the input can be masked with the context being sufficient to predict the blanks (Ericsson et al. 2022b). Equivariance and invariance of pretext and downstream task should be carefully aligned, e.g. for classification, invariance to spatial orientation is beneficial, while it is counter-productive for object location (Ericsson et al. 2022b). Still, evaluating the exact effect of the pretext task is impaired by the "difficulty in making objective and direct comparisons between methods" (Morningstar et al. 2024). In the following chapters, the complexity, representation geometry, the influence of the pretext task on the downstream task and the benefits of multiple pretext tasks are examined in more detail.

6.1 Complexity of the pretext task

The complexity of the pretext task is determined by how difficult the task is to solve, which in turn influences the downstream performance and the learned representation quality (Goyal et al. 2019). Complex pretraining is especially beneficial on large architectures as it ensures that these architectures do not overfit on the downstream dataset (Goyal et al. 2021). An appropriate degree of complexity can also prevent short-cut learning (Robinson et al. 2021b). Transformations induced in the pretext task should not produce easily detectable artifacts. According to Haja et al. (2023), complexer models outperform their simpler counterparts, although the effect size is small. The variety of learned features increases, but they do not necessarily become more complex (Haja et al. 2023). While the complexity depends on the specific pretext design and is generally difficult to compare, the ratio of missing information and the extend of spatial misalignment can be indications (Tian et al. 2022). Strategies to adjust complexity can vary greatly. CL with negative samples and InfoNCE can be modified by creating views which are easier or more difficult to distinguish (Robinson et al. 2021b). This can be achieved by choosing harder negative sampling strategies and adjusting the temperature in the InfoNCE loss term (Robinson et al. 2021b). For Jigsaw, the number of permutations can be increased to create a harder task, while for Colorization, the parameter k in the kNN calculation can be elevated (Goyal et al. 2019). Pretext tasks vary in sensitivity regarding pretext complexity: a more complex colorization task e.g. improves the results less significantly compared to a more complex jigsaw task (Goyal et al. 2019). Kim et al. (2018) complicate Jigsaw, Inpainting, and Colorization achieving higher accuracies for both classification (up to 1.2 percentage points improvement) and segmentation (up to 1.1 percantage points) using PASCAL VOC 2007, indicating that more difficult pretext tasks result in more useful representations. Tian et al. (2022) compare the influence of the difficulty regarding the pretext tasks zoom-in, zoom-out, image distortion, blurring and grayscale with ViT and ImageNet in a generative and token-based setting. All tasks prefer medium difficulty (Tian et al. 2022). Feng et al. (2020) reconstruct the whole image from a cropped part and conclude that oversized crops result in an inability to learn the semantics of the image, which causes trivial solutions, while the images contain too little information to recover the whole sample if the crop is too small. For CL, the difficulty of the discrimination task does not only determine the quality of the feature extraction, but also which features

are learned (Robinson et al. 2021b). Complex pretraining is especially beneficial on large architectures and large datasets, also to avoid overfitting (Goyal et al. 2021) and to enable learning useful representations. On the other hand, too hard and specific pretext tasks might lead to learning task-specific details that do not generalize well (Jing and Tian 2020). Thus, although the appropriate difficulty or complexity of the pretext task is an important factor, the correct level of difficulty must be determined individually in relation to the model size and the downstream task.

6.2 Representation geometry

The intrinsic dimensionality of the pretext representation can be understood as a tuneable hyperparameter (Wallace and Hariharan 2020). Useful representations should be expressive and reasonably sized (Bengio et al. 2013; Kumar et al. 2022b). The representation dimensionality should be sufficiently large, lower-dimensional sub-spaces should be avoided (Dubois et al. 2022). The representation geometry is not tied to the respective SSL type, it can vary significantly within one framework (Cosentino et al. 2022b). While MoCo-v1 and MoCo-v2 vary significantly in equivariance spread, PCL and SimCLR versions have a similar geometric representation (Cosentino et al. 2022b). Pretext tasks can vary across domains and the pretext task complexity does not necessarily determine the compactness of the learned representation (Wallace and Hariharan 2020). This might be one reason why harder pretext tasks do not necessarily result in a better downstream performance. Rotation has a similar implicit dimensionality as the significantly more complex instance-discrimination, while auto-enconding and Jigsaw have more compact representations (Wallace and Hariharan 2020). More complex pretext tasks might not necessarily make use of the available latent space (Wallace and Hariharan 2020). In turn, suitable data augmentations strategies can boost the representation quality of CL (Jaiswal et al. 2020; Misra and Maaten 2020). According to Dubois et al. (2024), increasing the representation dimensionality improves linear separability of the representation as the linear classifier is dependant on the input dimension. At the same time, this can decreases probe generalization due to overfitting (Dubois et al. 2024). According to Cosentino et al. (2022b), it is possible to "predict the transfer learning capability for a specific model based on the geometric properties of its semantic and augmentation manifolds". They consequently link representation geometry and performance (Cosentino et al. 2022b). According to Gupta (2025), downstream tasks such as encoding relations of objects require rich structural relationships that cannot be captured by learning similarities. Instead, minimizing an equivariance objective can improve downstream performance (Gupta 2025). For downstream data similar to ImageNet, the performance is negatively correlated with geometric properties such as semantic equivariance, spread of augmentation equivariance, rotation and augmentation affinity (Cosentino et al. 2022b). Instead, linearization and invariance are key indicators of transferability (Cosentino et al. 2022b). Representations which are covariant to the pretext transformation are helpful for predicting 3D correspondences, but counter-productive regarding most semantic recognition tasks, where invariant representations should be preferred (Ji et al. 2019; Misra and Maaten 2020). Covariant transformations are favored by pretext tasks predicting a property of the pretext image transformation (Misra and Maaten 2020). Transfer performance is correlated with higher-order statistics of the backbone encoder when the downstream dataset is more different from ImageNet (Cosentino et al. 2022b). The pretrained encoder representa-

tions are more meaningful than those of the decoder, the latter come with a risk of overfitting inductive information (Zhang et al. 2023c).

6.3 Influence of pretext on downstream performance

The complex relationship of pretext and downstream task is a key to successful SSL (Wallace and Hariharan 2020) and makes the effective pretext task design one of the most important challenges in SSL (Jing and Tian 2020). Obviously, the pretext task should learn features relevant to the downstream task (Yamaguchi et al. 2021) and its objective should be sufficiently similar to the downstream objective to be able to subsume it (Mao 2020; Sehwag et al. 2019). Although the interaction of pretext and downstream task is so important to SSL, the exact influence of the pretext task has hardly been studied. Most work focuses on improving the pretext task design by measuring downstream improvements on benchmarking data, while it is questionable how well this evaluation measures pretext representation quality or downstream performance on other datasets (Abnar et al. 2021; Marks et al. 2025). Is the pretraining performance of different pretext tasks an indicator for their downstream performance? Do pretraining and downstream performance correlate? What are helpful metrics to evaluate the pretext task? How generalizable is the pretext task across different downstream tasks? These questions are discussed in the following sections.

6.3.1 Correlation of pretext and downstream performance

Xie et al. (2023d) analyze different Swin-T-v2 models in combination with SimMIM as well as ImageNet-1K and ImageNet-22K for pretraining and FT on ImageNet-1K, iNaturalist-18, MS COCO and ADE20K as downstream task. The correlation between training loss and downstream accuracy is negative in overfitting and positive in non-overfitting regimes (Xie et al. 2023d). El-Nouby et al. (2024) report a positive correlation of pretext DFN-2B validation loss and downstream Imagenet-1K top-1 validation accuracy with AIM on ViT. According to Yi et al. (2022), the loss of simple pretext tasks such as rotation correlates with the downstream classification loss for CIFAR10, Caltech and Imagenet. Unfortunately, this does not work for pretext tasks using contrastive loss, supposedly due to the class bias of contrastive loss and the strong influence of augmentations (Yi et al. 2022). According to Marks et al. (2025), the pretext performance can be used for ranking, but not to estimate the absolute performance. Additionally, pretext task performance is not the only influencing factor, as larger models perform more effectively at a given pretext validation loss value (Chen et al. 2020d). Abnar et al. (2021) analyze the influence of the pretext on the downstream task using eight different pretraining datasets and JFT-300 for downstream training. Increasing training duration and model size as well as dataset size increases pretext performance (Abnar et al. 2021). Still, the pretext accuracy has a stronger predictive power than the other factors, so the authors argue that the effect of the model size and pretext task data size is only indirect via the pretext accuracy (Abnar et al. 2021).

Pretext and downstream performance are not always positively correlated (Gui et al. 2024; Zhang et al. 2023c). SSL comes with a trade-off between learning invariant features useful for the pretext and for the downstream task: While pretext "representations with higher class separation" perform better on the pretext task, their features are less valuable for the downstream task (Kornblith et al. 2021). Pretraining in general can suffer from a

pretext category bias when only information related to the pretext task and its classes is learned (El-Nouby et al. 2021). In such cases, a higher pretext performances can even lead to worse downstream task performance (Tschannen et al. 2020). In turn, a higher pretext loss can result in a better downstream generalization, especially given harder pretext tasks (Ryali et al. 2021), as the pretext loss does not always behave monotonically in relation to the generalization ability (He et al. 2020; Kolesnikov et al. 2019; Ryali et al. 2021). Robinson et al. (2021b) train encoders with ResNet-18 backbones using SimCLR and report that a lower InfoNCE loss results in lower errors for MNIST and Trifeature color downstream tasks, while the correlation is negative regarding Trifeature shape, texture and STL10, indicating that this pretext task only improves learning some features.

Arora et al. (2019) discuss the problem for CL from a theoretical perspective: they consider semantic similarity as latent classes, where similar pairs are sampled from the same latent class, while the downstream task consists of a subset of the latent classes. With a large number of latent classes, a small pretext loss should also result in a small (supervised) downstream loss (Arora et al. 2019). The pretext loss increases if positive and negative views come from different latent classes or if the intra-class deviation is high (Arora et al. 2019). Abnar et al. (2021) pretrain 2916 different models (1500 on ViT, 1400 on MLP-Mixers, 16 on ResNets) on JFT-300 dataset for few-shot downstream tasks and discover a power-law relation between the two tasks. The downstream accuracy saturates considerably below 100% given a wide range of shots and pretext tasks (Abnar et al. 2021). This saturation value is dependent on the Bayes error (an intrinsic error of the task definition), the amount of available samples and alignment of pretext and downstream task (Abnar et al. 2021). The pretext-downstream performance curve shape also varies with dataset: Oxford-IIIT-Pets, Cars, Caltech101, DTD, and UC Merced are rather characterized by a saturation curve, while Cifar100 and ImageNet are more similar to a linear relation and Colorectal Histolog is nearly horizontal (Abnar et al. 2021). Still, the pretext task ranking is consistent over different downstream tasks (Abnar et al. 2021).

These findings demonstrate that a high-performing pretext task does not guarantee a good downstream performance due to overfitting to the pretext objective and the saturation phenomenon. Furthermore, the pretext-downstream task relation varies with dataset, so a specific pretext task performance cannot be evaluated without taking the dataset and the downstream task into consideration.

6.3.2 Alternative evaluation metrics

The lack of a pretext-downstream performance correlation calls for another reliable evaluation metric. Possibilities are e.g. the effective rank of the SSL representation (Garrido et al. 2023a), intra-class data continuity (HaoChen et al. 2022a), conditional independence (CI) between the pretext labels and the downstream data (Zaiem et al. 2022), MI (Tian et al. 2020a) or the extent of pretext feature reuse (measured via Centered Kernel Alignment) (Neyshabur et al. 2021; Zhang et al. 2023c). Tian et al. (2020a) analyzed the relation between downstream performance and MI: When augmentations that manipulate the color-space are applied, minimal MI results in the best LP accuracy on STL10. If the representations are learned using patches with different offsets from each other, very small or very high MI values degrade the performance (Tian et al. 2020a). In CL, views should not share all information as this discards nuisances, while no shared information withdraws the incen-

tive for learning (Tian et al. 2020a). Low-dimensional SSL representations learned under the CI assumption "can express the ground truth label as a linear function" (Lee et al. 2021). The bounding function decreases quadratically with CI and linearly with the target data size (Zaiem et al. 2022). Still, the assumption of linear modeling could be very limiting for complex downstream tasks (Zaiem et al. 2022) and augmented samples are often strongly correlated (HaoChen et al. 2022a), so CI is far from a perfect metric and the CI condition is rarely satisfied in practice (Ericsson et al. 2022b; Teng et al. 2022). Dubois et al. (2024) propose a scaling law based on four sources of error (approximation, representation usability, probe generalization and encoder generalization error) which is tested on 169 models. A small generalization error seems to be a good predictor (Dubois et al. 2024). More powerful pretext tasks also result in a higher confidence downstream calibration, meaning that wrong predictions are identifiable by low-confidence probabilities (Ericsson et al. 2021; Guo et al. 2017). An improved pretext performance also correlates with larger attention areas, which can improve the transfer to recognition tasks (Ericsson et al. 2021).

6.3.3 Pretext generalization

Ideally, a pretext task should find features in every pretraining dataset that are useful to a wide range of downstream tasks. But how realistic is the ideal of a universal pretext task? While there are some cases of pretext tasks being superior across different types of downstream tasks and datasets (Abnar et al. 2021; Jing and Tian 2020), most SSL methods for CV are not optimized concerning transferability, especially if the downstream data is too different from pretext data (Ericsson et al. 2021, 2022b; Goyal et al. 2019; Van Horn et al. 2021). In contrast, in NLP broader benchmarks are used and transferability is more in focus (Brown et al. 2020; Devlin et al. 2018; Ericsson et al. 2022b). For CV, pretext-ranking varies across downstream tasks, architectures and hyperparameter settings (Abnar et al. 2021; Newell and Deng 2020). Ericsson et al. (2021) compare 13 self-supervised models on 40 downstream tasks with diverse datasets and conclude that there is no universal pretext task. The pretext performance can vary significantly over different downstream task types: while PCL-v1 performs best for semantic segmentation, it performs worse than all other analyzed frameworks for linear many-shot recognition and few-shot recognition (Ericsson et al. 2022b; Li et al. 2021e). Even very similar pretext tasks can have considerably different results across downstream task types: PCL-v2 has an average performance of 56.73% across 11 datasets for linear many-shot recognition, while PCL-v1 has 69.13% (Ericsson et al. 2021). For semantic segmentation, both nearly have the same performance (Ericsson et al. 2021). Furthermore, different downstream task types require different pretext task characteristic as they rely on the exploitation of different data aspects (Zaiem et al. 2022). Consequently, pretext tasks suitable for downstream classification cannot automatically be generalized to other applications (Newell and Deng 2020). Still, several strategies to improve generalization have been proposed. Saito et al. (2020) suggest to use self-supervised neighborhood clustering with entropy optimization to grasp the structure of the target domain instead of trying guessing the "category overlap" of the pretraining and the target domain for universal domain adaption. Automatic pretext design (Wang et al. 2021; Xie et al. 2021b), universal sub-networks (Chen et al. 2021a), processing 2D as well as 3D images by including "switchable patch embedding" (Xie et al. 2022c), or combining multiple pretext tasks (Wu et al. 2021; Maninis et al. 2025) could be beneficial.

6.4 Multiple pretext tasks

Several authors have combined multiple pretext tasks to boost performance (Chen et al. 2020b) and generalization (Wu et al. 2021). The learned representations should have more representative features due to more diverse supervision, Jing and Tian (2020) even recommend to use multiple pretext tasks "whenever possible". Combinations are especially relevant for predictive SSL to increase difficulty and enable successful representation learning. Still, generative (e.g. (Tian et al. 2022)) and contrastive frameworks (e.g. PIRL (Misra and Maaten 2020)) can also benefit from combinations. Table 9 summarizes pretext task combinations tested on ImageNet using self-pretraining. Nearly all combinations are more beneficial than just using one pretext task. Monocular depth estimation with NYUv2 and KNN classification on ImageNet are the only case where the addition of a pretext task (MIM) decreases the performance, although the decrease for ImageNet is very small (Maninis et al. 2025). The improvements of the other examples vary: on the two MIM-based approaches, only minor improvements were achieved in contrast to contrastive PIRL and the predictive approaches. PIRL with Rotation achieves 60.2% top-1 accuracy on Imagenet and 62.2% with Jigsaw; but if Rotation and Jigsaw are combined, the accuracy is 63.1% (Misra and Maaten 2020). The tendency is similar for PASCAL VOC 2007, Places205 and iNaturalist (Misra and Maaten 2020). Sole Jigsaw achieves 66.6% on classification of PASCAL VOC and 36.8% on segmentation, but the combination of Jigsaw, Inpainting and Colorization reaches 69.2% and 39.3% (Kim et al. 2018). The combination of spatial masking and spatial misalignment improves pretraining with BEiT and ViT-Base/16 (Tian et al. 2022). Tian et al. (2022) conclude that MIM pretraining profits mostly from missing information, but including spatial transformation adds synergies. Zhang et al. (2024b) combine Relative Orientation Regression and Relative Location Regression. Their pretext task MLT shows considerable improvements over SimCLR, BYOL, MAE, Jigsaw, Model Genesis, and Relative Location for ACDC CMR and Knee MRI classification given ID pretraining (Zhang et al. 2024b). The improvements are more considerable if the downstream dataset is smaller (Zhang et al. 2024b). Ji et al. (2022) combine different SSL paradigms and stack pretext task, including semantic class prediction, rotation, and contrastive prediction. Their multi-task model is superior to a single pretext task in all analyzed cases where the backbone depth matches the dataset and the model is therefore able to optimize three tasks simultaneously (Ji et al. 2022).

Combining pretext tasks raises the question whether and how these interact and how to design an ideal combination. According to Zaiem et al. (2022), one key aspect for multi-pretext models is understanding the interaction between the different pretext task labels. The pretext tasks can either be learned sequentially (e.g. with increasing difficulty (Shi et al. 2023a)) or via joint optimization (Ma and Liu 2023). The sum of losses can be minimized together or the combination can be regarded as a single task with a joint distribution (Lee et al. 2020). There are different possibilities to balance multiple pretext tasks, the trade-off has a significant influence on the final result (Ma and Liu 2023). Tasks can e.g. be weighed according to the reciprocal of the average validation loss (Wu et al. 2021), their task-dependent uncertainty (Kendall et al. 2018; Ma and Liu 2023), the attention mechanism (Li et al. 2021c) or estimated sparse weighting vectors based on CI (Zaiem et al. 2022). Additional difficulties are that training can end in pareto-optimal states, which causes early

Table 9 Performance improvements (Gain) of different SSL methods with different architectures due to combining pretext tasks for ImageNet pretraining and different downstream datasets (Down data). The performance metrics are accuracy (Acc), AUC, mIoU, and Root-MSE (RMSE). The gain is calculated in relation to the previous combination(s). PIRL results originate from Misra and Maaten (2020), Rel. Position from Doersch and Zisserman (2017), BEiT from Tian et al. (2022), MAE from Ahn et al. (2022), MTL from Zhang et al. (2024b), CDJP from Kim et al. (2018) and TWIST from Maninis et al. (2025)

Method	Architecture	Pretext task(s)	KT	Down data	Metric	Performance	Gain
PIRL	ResNet-50	Rotation	LP	ImageNet	Acc	60.2	–
PIRL	ResNet-50	Jigsaw	LP	ImageNet	Acc	62.2	–
PIRL	ResNet-50	Rotation + Jigsaw	LP	ImageNet	Acc	63.1	3.2
Rel. Position	ResNet-101-v2	Rel. Position	LP	ImageNet	Acc	59.2	–
Rel. Position	ResNet-101-v2	Rel. Position + Colorization	LP	ImageNet	Acc	66.6	12.5
Rel. Position	ResNet-101-v2	Rel. Position + Exemplar	LP	ImageNet	Acc	65.2	10.2
Rel. Position	ResNet-101-v2	Rel. Position + Motion Segmentation	LP	ImageNet	Acc	63.7	7.6
Rel. Position	ResNet-101-v2	Rel. Position + Colorization + Exemplar	LP	ImageNet	Acc	68.7	3.0
Rel. Position	ResNet-101-v2	Rel. Pos. + Color. + Exemplar + Motion Seg	LP	ImageNet	Acc	69.3	6.2
BEiT	ViT-B/16	Masking	FT	ImageNet	Acc	82.9	–
BEiT	ViT-B/16	Zoom-in	FT	ImageNet	Acc	82.7	–
BEiT	ViT-B/16	Masking + Zoom-in	FT	ImageNet	Acc	83.2	0.5
MAE	ViT-Tiny	Masking	FT	ImageNet	Acc	72.6	–
MAE	ViT-Tiny	Masking + Position Prediction	FT	ImageNet	Acc	72.7	0.1
MTL	U-Net	Relative Location Regression	FT	ACDC CMR	AUC	0.962	–
MTL	U-Net	Relative Orientation Regression	FT	ACDC CMR	AUC	0.934	–
MTL	U-Net	Rel. Location Reg. + Rel. Orientation Reg	FT	ACDC CMR	AUC	0.972	2.6
CDJP	Siam. AlexNet	Jigsaw	FT	PASCAL VOC	Acc	66.6	–
CDJP	Siam. AlexNet	Jigsaw + Inpainting	FT	PASCAL VOC	Acc	67.4	1.2
CDJP	Siam. AlexNet	Jigsaw + Colorization	FT	PASCAL VOC	Acc	68.4	2.7
CDJP	Siam. AlexNet	Jigsaw + Inpainting + Colorization	FT	PASCAL VOC	Acc	69.2	2.9
TIPS	ViT-B	CLIP + self-distillation	LP	PASCAL VOC	mIoU	70.3	–
TIPS	ViT-B	CLIP + self-distillation + MIM	LP	PASCAL VOC	mIoU	75.9	8.0
TIPS	ViT-B	CLIP + self-distillation	LP	NYUv2	RMSE	0.589	–
TIPS	ViT-B	CLIP + self-distillation + MIM	LP	NYUv2	RMSE	0.511	**−13.2**
TIPS	ViT-B	CLIP + self-distillation	KNN	ImageNet	Acc	79.1	–
TIPS	ViT-B	CLIP + self-distillation + MIM	KNN	ImageNet	Acc	79.0	**−0.1**

stopping (Shi et al. 2023a) and the required computational resources (Noroozi et al. 2018). To date, most publications focus on a single pretext task.

7 Network size

The tendency of larger networks performing better is more pronounced for SSL compared to SL due to the additional information that must be absorbed, especially for FT (Chen et al. 2020b; He et al. 2016; Kolesnikov et al. 2020; Mao 2020). Scaling up pretext training can both improve feature quality and generalization (El-Nouby et al. 2021; Hu et al. 2025; Khan et al. 2025b; Rani et al. 2023). Bigger networks are especially useful in combination with challenging pretext tasks and large pretext datasets (Chen et al. 2020c; Ericsson et al. 2022b; Devlin et al. 2018; Goyal et al. 2019; Jing and Tian 2020). Larger SSL models are characterized by higher sample efficiencies and require less optimization steps (Xie et al. 2023d). According to Spencer et al. (2022), older SSL frameworks can outperform newer systems when simply updating the backbone on monocular reconstruction (e.g. ConvNeXt-B instead of ResNet-18), demonstrating the great impact of the architecture and size. They achieve relative improvements of up to 25% on 16 benchmarking datasets (Spencer et al. 2022). Downstream performance often profits from higher pretraining network capacities, while the downstream model itself can be smaller (Chen et al. 2020c; Ericsson et al. 2022b; Goyal et al. 2019; Newell and Deng 2020). Decoupling the size of the pretext and downstream model can be a useful if both tasks have significantly different size requirements.

In general, both increasing network width and depth has a positive impact (Misra and Maaten 2020; Noroozi et al. 2018; Zeng and Xie 2020; Zhang et al. 2021c; Zhou et al. 2021). As shown in Fig. 7, all models increase LP top-1 ImageNet accuracy with higher widths of ResNet-50, but the improvements differ significantly and the increase slows down given bigger network sizes (Chen et al. 2020c). Augmenting the width from ResNet-50 to ResNet-50 (2x) boosts performance on average by 5.1%, increasing from ResNet-50 (2x) to ResNet-50 (4x) only results in 2.2% average improvement, according to the data from Fig. 7. VicReg is characterized by the lowest average improvement (1.6%), MoCo by the highest (6.4%). MoCo has the worst absolute performance, SwAV and BYOL have the best. Due to the different sensitivity regarding network width, the SSL framwework that performs best on ResNet-50 is not necessarily the best model on ResNet-50 (4x).

Fig. 8 demonstrates the influence of the network size on self-supervised ViT, either using FT (Fig. 8a) or LP (Fig. 8b). BeiT on ViT-H with LP, where the performance is 0.1 percentage points lower compared to ViT-L, is the only case where a bigger networks does not improve the performance. As CL tends to be better regarding LP and MIM regarding FT and many authors only test their models on the superior metric, Fig. 8a includes more MIM models and Fig. 8b includes more CL ones. The models that perform best on FT (Beit-v2, PeCo, CAE-v2, and data2vec) are all generative and the average FT-performance of generative models is 2.9 percentage points higher compared to contrastive models. In turn, the top LP models include two contrastive ones (Mugs and DINO), three generative ones (CAE-v2, iBot, MSN) and one hybrid model (MAE-CT). The results for FT (80.6%−88.3%) are much closer to each other compared to LP (51.8%−82.2%). LP seems to be significantly more sensitive regarding small network sizes. The average performance increase when scaling up from ViT-S to ViT-B is 6.7% for LP and 2.0% for FT. For the increase from ViT-B to

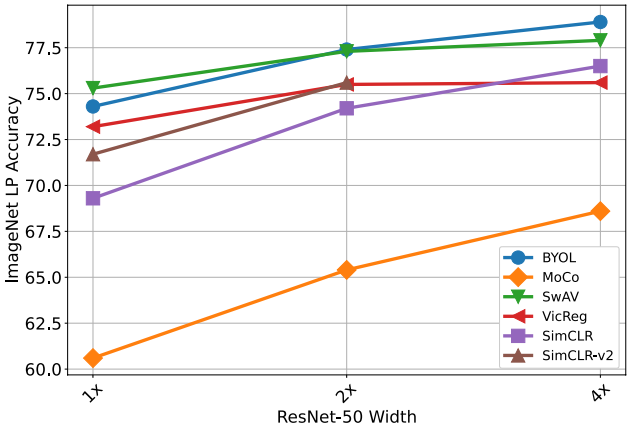

Fig. 7 Impact of varying ResNet-50 width on ImageNet top-1 downstream accuracy with LP and contrastive SSL according to the original publications (Bardes et al. 2021; Caron et al. 2021a; Chen et al. 2020b, c; Grill et al. 2020; He et al. 2020) and Ozbulak et al. (2023)

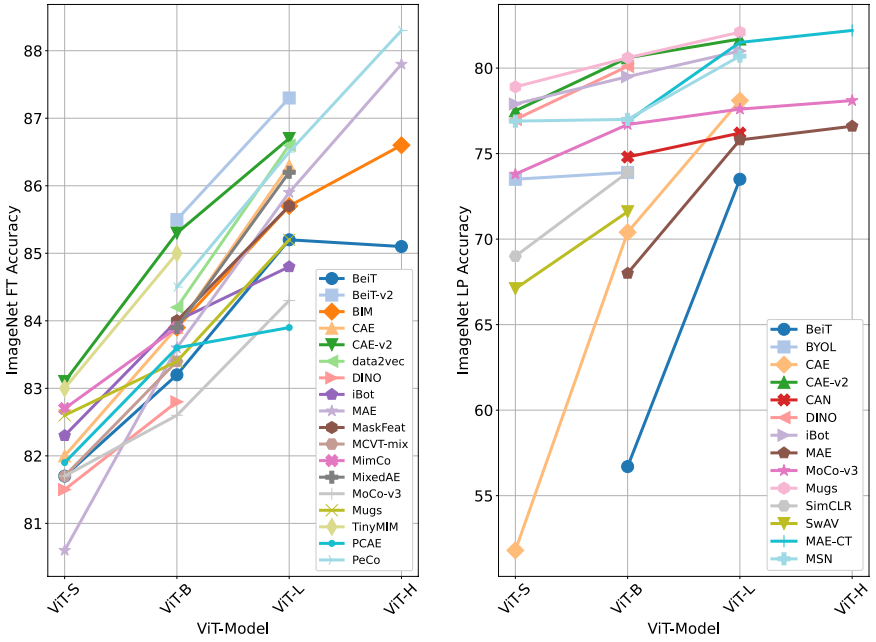

Fig. 8 Impact of ViT size on ImageNet top-1 accuracy with (a) FT top-1 ImageNet accuracy on different ViT architectures derived from the corresponding original publications (Baevski et al. 2022; Bao et al. 2021; Caron et al. 2021b; Chen et al. 2021c, 2023c, 2024a; Dong et al. 2023a; He et al. 2022; Li et al. 2022c; Mo et al. 2023; Peng et al. 2022; Wei et al. 2022a; Zhang et al. 2023e; Zhou et al. 2022a, b, 2023c) and Ozbulak et al. (2023) and (b) LP top-1 ImageNet downstream accuracy on different ViT architectures derived from the corresponding original publications (Assran et al. 2022b; Bao et al. 2021; Caron et al. 2021a, b; Chen et al. 2020b, 2021c; Grill et al. 2020; He et al. 2022; Lehner et al. 2024; Mishra et al. 2022; Zhang et al. 2023e; Zhou et al. 2022a, b) as well as Ozbulak et al. (2023).

ViT-L, the difference is reduced (4.6% vs. 2.1%), for the increase from ViT-L to ViT-H, the average improvement of FT is even slightly higher (0.9% LP vs. 1.3% FT). For LP, it is therefore particularly recommendable using at least ViT-B or ViT-L. Some models, e.g. PeCo, CAE-v2, and TinyMIM, do not change ranking no matter which ViT model is used, indicating that the choice of the pretext task is more influential than the choice of network size in some cases.

Different learning paradigms have different size requirements: In general, CL outperforms MIM given small ViT-networks, for large models MIM is more effective (Park et al. 2023; Wang et al. 2023d). MIM tends to require larger or more sophisticated models to manage higher degrees of complexity, especially if high masking ratios are applied (Filiot et al. 2023; Mao et al. 2022) to prevent trivial solutions (Wu et al. 2025). Furthermore, MIM is more sample efficient and can handle higher network capacities better given the same pretraining data (Moutakanni et al. 2024). Zhou et al. (2022c) suppose that reconstruction-based pretext tasks result in more effective pretraining epochs than CL, consequently they can learn more in the same number of epochs and suffer less from overfitting given large models and long pretraining. Additionally, CL can require huge computational resources compared to MIM, where example interaction is limited, so scaling up MIM-models is easier (Gui et al. 2024). "Information encoded by the contrastive loss is limited", so CL can degrade when complex models are trained on smaller datasets (Cao and Wu 2021). Still, CL requires networks having a sufficiently high capacity to learn to discriminate between instances (Caron et al. 2021a; Chen et al. 2020b, c; Goyal et al. 2019; Grill et al. 2020) and MIM does not always benefit from large networks (Filiot et al. 2023). Filiot et al. (2023) test the scalability of MIM on different ViT sizes and histopathology data: while scaling from ViT-S to ViT-B results in an average increase of of 2.5% across all histopathology tasks, further enlarging to ViT-L does not come with clear improvements. In general, large models are hard to handle and computationally expensive (Ohri and Kumar 2021), performance can plateau or decrease given an excess in pretraining capacity (Cabannes et al. 2023; El-Nouby et al. 2024; Grill et al. 2020) and computational costs can grow faster than performance (Fernandez et al. 2025). Consequently, while larger networks lead to major improvements, it seems to be of equal importance that the dataset size matches the network size, the pretext task complexity, the KT type, as well as the data properties and the downstream task. Also, most of the demonstrated results use ImageNet. Girish et al. (2022) correlate the number of parameters and top-1 accuracy on different datasets using various variants of ResNet and MobileNet that have been pretrained with SimCLR and ImageNet. While large SSL models perform well on ImageNet (Caron et al. 2021a; Chen et al. 2020c, b; He et al. 2020), this is not the case for downstream tasks being significantly different (Girish et al. 2022). For ResNet-like architectures, the downstream performance on the Aircraft is better using lighter networks while for most datasets on MobileNet-like architectures, the relationships are highly non-linear and non-monotonic (Girish et al. 2022). This can be explained by architecture recipes being not transferable to other datasets (Girish et al. 2022). Therefore it important to note that no universal scaling laws seems to exist; only general trends can be observed.

8 Influence of the pretext data

Self-supervised pretraining can be performed either with the downstream dataset itself (self-pretraining) or with an alternative (un)labeled dataset, which often is considerably larger (Zhou et al. 2023a). If different pretraining and downstream datasets are are used, there are two options: either a dataset from another domain is used for OOD pretraining (e.g. pretraining on natural images and downstream training on CT scans), or ID pretraining is possible if a pretraining dataset from the same domain is available. OOD-pretraining on a standard benchmarking data is common, but can introduce considerable discrepancies if the available data is too different (El-Nouby et al. 2021; Kawashti et al. 2023; Krishna et al. 2023). What data properties should the ideal pretext dataset possess (Sect. 8.1)? How large should it be (Sect. 8.3)? Is ID pretraining important (Sect. 8.4)? Is self-pretraining beneficial (Sect. 8.5)? How representative are results obtained with the common benchmarking datasets ImageNet (Sect. 8.2)? These questions and their implications for the pretext task design are analyzed in detail in this chapter. An overview of important datasets for SSL including their sizes and domains can be found in Table 10.

8.1 Dataset characteristics and quality

Pretraining is only beneficial if the pretext dataset properties and the design of the pretext task match with the downstream task and result in semantic knowledge gain (Su et al. 2020). The properties of the downstream dataset have to be taken into account when designing or choosing an appropriate pretext task (Zhang et al. 2023c). Since the pretext task strives to exploit the pretext data structure, its efficiency is dataset-dependent (Ericsson et al. 2022b) and the relation between the pretraining and downstream data distributions is of great importance (Cole et al. 2022). Consequently, one SSL method does not dominate all others across datasets and domains (Huang et al. 2024a). Well-balanced and diverse distributions improve both performance and generalization if the pretext-downstream data distribution distance is low (Al Kader Hammoud et al. 2025; Li et al. 2021d). In contrast, if downstream and pretext distribution differ significantly, a high pretext data diversity can decrease downstream performance (Al Kader Hammoud et al. 2025). Dataset characteristics which can induce shortcut learning should be avoided, such as color aberrations, watermarks or other exploitable low-level visual features (Minderer et al. 2020). The pretext dataset has to be in accordance with the pretraining objective, e.g. rotation captures pose well, but struggles to distinguish color and textures or rotation-invariant data (Wallace and Hariharan 2020; Yamaguchi et al. 2021). Instead, rotation requires canonical orientation and has difficulties with highly symmetrical images (Su et al. 2020). Jigsaw needs objects of recognizable shapes and is not suitable for classification tasks requiring fine-grained distinctions in color or texture (Wallace and Hariharan 2020). The performance of predictive tasks based on spatial information also depends on the image size, as bigger images result in a higher pretext accuracy, but using comparable sizes for pretext and downstream task is a good approach (Taleb et al. 2000; Zhang et al. 2023c; Zhuang et al. 2019a). The image size "determines the amount of spatial information from which the network extracts features", which is especially relevant for segmentation (Zhang et al. 2023c). Reconstruction-based SSL can decrease in accuracy with color, texture, viewpoint, and lighting variations, probably because considerably more details have to be encoded which are not beneficial for the

Table 10 Overview of important datasets including the respective sizes and domains

Dataset	Size	Domain
3D CT (Jiang and Veeraraghavan 2024)	10,412	Medical
ACDC CMR (Bernard et al. 2018)	100	Medical
ADE20K (Zhou et al. 2019b)	27,574	Scenes
Aircraft (Maji et al. 2013)	10,200	Vehicles
BACH (Aresta et al. 2019)	400	Medical
BigEarth20k (Sumbul et al. 2019)	20,000	Geographical
BigEarthNet (Sumbul et al. 2019)	590,326	Geographical
Caltech101 (Li et al. 2022a)	9,146	General
Caltech256 (Griffin et al. 2007)	30,608	General
Cars (Krause et al. 2013)	16,185	Vehicles
Cashew1k (Yin et al. 2023)	1,750	Geographical
ChestX-ray14 (Wang et al. 2017)	112,120	Medical
ChestX-ray8 (Wang et al. 2017)	108,948	Medical
CheXpert (Irvin et al. 2019)	224,316	Medical
Cifar-10 (Krizhevsky 2009)	60,000	General
Cifar-100 (Krizhevsky 2009)	60,000	General
CityScapes (Cordts et al. 2016)	25,000	Scenes
CLEVR (Johnson et al. 2017)	100,000	Medical
Colorectal Histology MNIST (Kather et al. 2016)	5,010	Medical
COVID-CT (Yang et al. 2020b)	349	Medical
CRC (Kather et al. 2018)	100,000	Medical
CUB200 (Wah et al. 2011)	11,788	Birds
Derm (Azizi et al. 2021; Liu et al. 2020)	20,676	Medical
DermaMNIST (Yang et al. 2021, 2023e)	10,015	Medical
Describable Textures (DTD) (Cimpoi et al. 2014)	5,640	Textures
DFN-2B (Fang et al. 2023a)	2,000,000,000	General
DSBCC (Newton et al. 2015)	1,140	Medical
Eurosat (Helber et al. 2019)	27,000	Geographical
Fast MRI Knee (Zbontar et al. 2019)	1,398	Medical
Flowers-102 (Nilsback and Zisserman 2008)	8,189	Plants
Food-101 (Bossard et al. 2014)	101,000	Food
FT-300 M (Sun et al. 2017)	300,000,000	General
GeoLifeCLEF 2020 (GLC20) (Cole et al. 2020)	1,900,000	Geo./ Bio
IG-3B (Singh et al. 2022, 2023)	3,000,000,000	General
ImageNet modified (Ogidi et al. 2023)	1,200,000	General
ImageNet-1K (Deng et al. 2009)	1,431,167	General
ImageNet-20 (Zhang et al. 2021d)	26,348	General
ImageNet-21K (Russakovsky et al. 2015)	14,197,122	General
ImageNet-21K-P (Ridnik et al. 2021)	12,358,688	General
iNat21-Plants (Ogidi et al. 2023)	1,100,000	Plants
iNaturalist-17 (Van Horn et al. 2018)	857,877	Biological
iNaturalist-21 (Van Horn et al. 2021)	3,286,843	Biological
NYUv2 (Silberman et al. 2012)	1,449	Scenes
JFT-300 M (Sun et al. 2017)	300,000,000	General
JFT-3B (Zhai et al. 2022a)	3,000,000,000	General
LC (Jiang and Veeraraghavan 2024)	196	Medical
LRad (Bakr et al. 2018)	139	Medical
LUNA16 (Setio et al. 2017)	888	Medical

Table 10 (continued)

Dataset	Size	Domain
MHIST (Wei et al. 2021)	3,152	Medical
Microsoft COCO (Lin et al. 2014)	328,124	Scenes
MMEarth (Nedungadi et al. 2024)	1,200,000	Geographical
MNIST (Deng 2012)	70,000	Digits
Montgomery-CXR (Jaeger et al. 2014)	138	Medical
Open Images (Kuznetsova et al. 2020)	1,910,098	General
OPPD (Leminen Madsen et al. 2020)	7,590	Plants
Oxford-IIIT-Pets (Parkhi et al. 2012)	7,349	Pets
PASCAL VOC 2007 (Everingham et al. 2007)	9,963	Scenes
PatchCamelyon (Veeling et al. 2018)	327,680	Medical
Places205 (Zhou et al. 2018)	2,500,000	Scenes
Places365 (Zhou et al. 2018)	1,811,528	Scenes
Pneumonia (Kermany et al. 2018)	5,863	Medical
Shenzhen Chest X-Ray (Jaeger et al. 2014)	662	Medical
Sport (Li and Fei-Fei 2007)	1,579	Sport
Stanford Knee MRI (Bien et al. 2018)	1,370	Medical
STL10 (Coates et al. 2011)	13,000	General
SUN-397 (Xiao et al. 2010)	108,753	Scenes
TCGA (Weinstein et al. 2013)	20,994	Medical
TCIA NSCLC (Aerts et al. 2015)	350	Medical
TFC (Field)/TFC5 (Beck et al. 2021)	540,000	Plants
TinyImageNet (Le and Yang 2015)	100,000	General
Trifeature (Hermann and Lampinen 2020)	100,000	Texture
TULIP (Kang et al. 2023)	15,672	Medical
UCM (Yang and Newsam 2010)	2,100	Geographical
Wukong (Gu et al. 2022)	101,483,885	General
YFCC100M (Thomee et al. 2016)	99,200,000	General

downstream task (Newell and Deng 2020). In general, the modeling assumptions have to be aligned with the data characteristics: If a clear division between the decision boundaries and regions with high data density is expected, problems arise e.g. if the dataset actually consists of strongly overlapping Cauchy distributions (Gui et al. 2024). Pretext tasks with the underlying assumption of semantic consistency are not feasible for non-iconic datasets which are not consisting of single-centric-objects, as two views can be very distant (Li et al. 2022f). Another important aspect to consider is image resolution: According to Cole et al. (2022), the top-1 accuracy on ImageNet using SimCLR and LP decreases with reduced pretext image resolution. Cao and Wu (2021) instead state that "fine-grained details in high resolution images may be unnecessary, or may even confuse the contrastive loss". Smaller resolutions furthermore allow bigger batch sizes, which is important for many CL frameworks with negative samples (Cao and Wu 2021).

8.2 ImageNet for pretraining

Most pretext tasks are evaluated on ImageNet to ensure comparability. The curated dataset centers on one object per image, consists of fine-grained, balanced classes and its distribution is less heavy-tailed compared to natural image collections (Li et al. 2022f; Tian

et al. 2021). An important question is whether ImageNet results can be generalized to other datasets, as this would avoid time-consuming analysis and allow practitioners to select the best model from an ImageNet-based comparison for their specific application. A higher difference between ImageNet and the downstream dataset makes pretraining with ImageNet less effective (Konstantakos et al. 2025). Experiments show that the more diverse content of uncurated images can be less easily exploited in pretext tasks tailored for curated datasets (Tian et al. 2021). For ResNet50 and BYOL, ImageNet top-1 accuracy decreases from 74.3% to 65.3% and 67.0% when pretrained on the uncurated datasets YFCC100M and JFT-300 M instead of ImageNet (Tian et al. 2021). The weights trained on ImageNet are supposed to focus on features helpful for diverse downstream dataset, so there should be no need of substantial weight changes during FT (Bengio et al. 2013; Matsoukas et al. 2022). This concept has been called into question by Raghu et al. (2019) who showed that instead of feature reuse, weight scaling and low-level statistics are the main reason for successful feature learning. While consistent and curated datasets enhance high-level visual representation learning, restricting pretraining to these may bias SSL development and limit downstream applications (Liu et al. 2021b).

CL frameworks, such as in MoCo, MoCo-v2, and SimCLR, do not actively ensure that the image category does not change due to data augmentations, they simply rely on the object-centric character of the images, especially regarding Crop (Purushwalkam and Gupta 2020). Consequently, the object-centric nature of ImageNet guarantees the latent prior of CL, namely that different views originate from the same object (Liu et al. 2021b). Liu et al. (2021b) even attribute the success of contrastive instance-level classification to benchmarking with ImageNet or similar datasets. According to Ericsson et al. (2021), representations can only be successfully transferred to other datasets with performance surpassing supervision in the many-shot regime using similar structured target datasets, the generalization of ImageNet results is lower for the few-shot regime. The transfer also works much better for recognition tasks compared to object detection and dense prediction (Ericsson et al. 2021). Girish et al. (2022) calculate rank correlation between top-1 accuracies on 11 downstream tasks obtained from ImageNet-pretrained ResNets and only report significant correlations for cases with target data very similar to ImageNet. Consequently, model rankings based on ImageNet can only generalize to similar datasets (e.g. Places205) (Kolesnikov et al. 2019), not to different domains (e.g. medical images (Raghu et al. 2019)). The utility of a transfer from ImageNet to images from the medical domain decreases with dataset size, domain gap as well as increased inductive bias (Matsoukas et al. 2022). Still, the transfer utility decreases to a lower extent with reduced model capacity (Matsoukas et al. 2022).

If large downstream datasets that have low similarities with ImageNet are used, ImageNet-pretraining is not recommendable (Matsoukas et al. 2022). Compared to more recent datasets such as Wukong (100 million image-text pairs), FT-300 M (300 million images) or JFT-3B (3 billion images), ImageNet (1 million images) is also rather small, which calls its general use for benchmarking into question. Pretext tasks might often overfit to ImageNet characteristics, e.g. by adjusting the image augmentations too much to ImageNet requirements (Van Horn et al. 2021). Consequently, research might focus on phenomena that only target ImageNet and similar datasets, but which are not necessarily relevant for less structured image datasets. This demonstrates the importance of a higher pretraining dataset diversity within SSL research.

8.3 Dataset size

While bigger pretraining datasets are likely to be beneficial for SL (Mensink et al. 2022; Hu et al. 2025; Xu et al. 2023), the scaling laws of SSL are yet to be uncovered (Lu et al. 2023). At the same time, pretraining on large datasets is the standard in SSL (El-Nouby et al. 2021; Reed et al. 2022), which has largely supported the success of SSL (Wang et al. 2023b). The pretext task can be effective given large datasets, even if they are noisy, unlabeled or scarcely labeled (Newell and Deng 2020). Large-scale pretraining almost always results in performance gains (Jonske et al. 2025), while the high capacity models often used for SSL tend to overfit given smaller datasets (El-Nouby et al. 2021). Still, in some cases large pretraining datasets can have negative effects, e.g. when pretraining a small model (Kolesnikov et al. 2020) or when larger pretraining datasets result in a more significant pretext data bias (Jaiswal et al. 2020). This could be caused by bigger datasets reducing the generalization of encoder and probe, which is shown by Dubois et al. (2024) with ImageNet-1K and ImageNet-22K. Also, pretraining can still be surprisingly good in case only small datasets are available (El-Nouby et al. 2021). Semantic equivariance, rotation affinity, augmentation affinity, and spread of augmentation equivariance even show a negative correlation with the dataset size (Xiao et al. 2021). Bigger downstream datasets also do not always profit from bigger pretraining datasets: BYOL classification on ImageNet-1K (1.8M) prefers the bigger JFT dataset (300 M), BYOL classification on Places-356 (1.8M) works better with the smaller JFT dataset (100 M) (Tian et al. 2021). According to Konstantakos et al. (2025), MAE and DINO on ViT-L are significantly more robust regarding low-data approaches compared to DeepCluster-v2 and SimCLR on ResNet-50. In the following chapters, the relationship between pretext dataset size and pretext paradigm (Sect. 8.3.1) as well as model size and pretext dataset size (Sect. 8.3.2) is discussed.

8.3.1 Dataset size and learning paradigm

The profitability of training MIM and CL with large datasets is a controversial topic with contradictory research results. Due to masking, MIM provides a huge amount of data samples itself, so scaling could be different from SL and other SSL approaches (Kaplan et al. 2020; Xie et al. 2023d; Zhai et al. 2022a). While MIM requires large datasets to scale up computation and model parameters, it does not benefit from more data in non-overfitting scenarios opposed to Masked Language Models (MLM) or supervised pretrained CV models (Xie et al. 2023d). Given SwinT-v2 in combination with SimMIM and sufficiently long training durations, there is a significant performance difference between large and small pretraining datasets (Xie et al. 2023d). Still, MAE seems to require less pretraining data compared to DINO "in order to learn generally applicable features that are inherently semantically discriminant and consistent with the visible objects" (Konstantakos et al. 2025). Lu et al. (2023) study the pretext scaling properties of self-supervised MIM with several downstream datasets (ImageNet, COCO, ADE20K and City Scapes). When the pretext dataset changes from ImageNet-1k to ImageNet-22k, which increases image quantity from 1 M to 14 M, there is a downstream performance gain with FT, especially on dense prediction tasks, but the performance gains saturate around 10 M images (Lu et al. 2023). Given LP, the pretraining size has a considerable influence if the pretraining domain differs from the downstream domain, so the saturation point increases to 100 M images (Lu et al. 2023). Consequently,

MIM can train large-capacity models with large datasets, but does not necessarily benefit from an excess of data (El-Nouby et al. 2021; Li et al. 2024a; Konstantakos et al. 2025; Wei et al. 2022b; Xie et al. 2023d). When being trained with minimal training data, MIM models can even yield performances comparable to those of models trained with larger datasets (El-Nouby et al. 2021). Consequently, MIM is an effective method to improve model capacity in low-data regimes (Lu et al. 2023).

While CL methods benefit from upscaling the network (El-Nouby et al. 2024; Oquab et al. 2024), they do not necessarily require large pretext datasets (Cole et al. 2022). Cole et al. (2022) show that SimCLR does only decrease in performance by 1 or 2% with 50% of ImageNet being available compared to the full dataset. Tian et al. (2021) use very large datasets instead of ImageNet without achieving considerable improvements, which could be attributed to a lack of experience with pretraining techniques for huge datasets (Wang et al. 2023b). Some CL frameworks also do not scale well with huge amounts of images, as each sample is considered as its own class and discriminating between them is expensive (Caron et al. 2021b). Furthermore, contrastive approaches need larger amounts of data when trained on uncurated datasets to reach similar performance (El-Nouby et al. 2021; Goyal et al. 2019). Another important aspect is the dimension of the learned representation with respect to the dataset size: larger and more complex representations are generally more useful for large datasets (Kolesnikov et al. 2019; Wallace and Hariharan 2020).

Table 11 analyzes the influence of the size of the pretraining dataset based on 29 combinations. In only four cases (14%), using a larger pretraining dataset reduces performance. In five cases, performance remains the same for the largest dataset compared to the second largest (17%). Otherwise, larger pretraining datasets also lead to better results (69%). For MIM, there seems to be a tendency that smaller networks profit less from large pretraining datasets than bigger ones. Comparing the result obtained with the largest available pretraining dataset with the results of the second largest, the average improvement is 0.7%. Comparing the largest and the smallest dataset, the average improvement is 1.5%. According to the results of Table 11, the contrastive models benefit more from bigger pretraining datasets compared to the generative models. However, most of the experiments were only carried out with a few dataset sizes, and the architectures and downstream datasets used differ greatly. More experimental data would be required for a more systematic evaluation.

8.3.2 Dataset size and model size

In SSL, the pretext task allows to extract more information per sample, which facilitates to use bigger and more data-hungry models with fewer samples compared to supervised models (Bao et al. 2021; Mao 2020). When pretraining from scratch on ImageNet, ViT-B is 1.4 percentage points better compared to ViT-L, which has 3.6 more parameters (Bao et al. 2021; Dosovitskiy et al. 2020). When applying SSL, the bigger model is 1.2 percentage points better, while the performance of both is significantly higher (Bao et al. 2021; Dosovitskiy et al. 2020). More efficient or harder pretext tasks might also benefit more from bigger datasets (Wang et al. 2023d): iBOT e.g. profits more from pretraining with ImageNet-22K instead of ImageNet-1K than BEiT (Zhou et al. 2022b). Still, SSL frameworks might not benefit from large pretraining datasets in combination with standard architectures (Cole et al. 2022).

Table 11 analyzes the impact of pretraining dataset and network size, indicating that the latter has a greater influence. When using SwinV2-G instead of SwinV2-B (1061 M instead of 87 M parameters) with SimMIM and Imagenet-1K for pretraining, top-1 accuracy is increased from 85.0% to 86.5% (Xie et al. 2023d). When using Imagenet-21K instead of Imagenet-1K, top-1 accuracy is only increased by 0.1 percentage points for SwinV2-B and there is no change with SwinV2-G (Xie et al. 2023d). With iNaturalist-18 as downstream dataset, the results are similar: the top-1 accuracy of SwinV2-L pretrained with ImageNet1K is 2.2 percentage points higher compared to SwinV2-B, while ImageNet-21K only achieves 0.1 percentage points more than ImageNet-1K on both SwinV2-B and SwinV2-L (Xie et al. 2023d). Increasing from ViT-B to ViT-L improves ImageNet-1K and ImageNet-22K accuracy by 1.8 and 2.3 percentage points for BEiT as well as 1.2 and 2.2 percentage points for iBOT (Zhou et al. 2022b). Increasing the pretraining dataset size from ImageNet-1K to ImageNet-22K only results in 0.3 (ViT-B) and 0.8 percentage points (ViT-L) gain for BEiT as well as 0.4 and 1.4 for iBOT (Zhou et al. 2022b). Consequently, ViT-L profits more from using the bigger dataset than ViT-B. In general, larger models seem to benefit more from bigger pretraining datasets (Baevski et al. 2023; Oquab et al. 2024; Singh et al. 2023) and bigger networks might be necessary to make use of large pretext datasets (Goyal et al. 2019). Additionally, increasing the pretraining dataset size is most effective given small datasets: Increasing from 10% ImageNet-1K to full ImageNet-1K increases accuracy by 1.9 percentage points on average, while increasing from ImageNet-1K to ImageNet-22K only yields 0.4.

When upscaling network and pretraining dataset size, saturation can occur. ViT model capacity saturates if the pretraining dataset is not upscaled in accordance with the architecture (Filiot et al. 2023). According to Table 11, increasing the pretraining size is always beneficial or neutral for architectures larger than ViT-B (86 M) and ResNet-50 (25.6M). Models pretrained with small dataset (6000 training samples or less) struggle to learn deeper layers, so smaller and more shallow architectures can be useful (Cao and Wu 2021; Xie et al. 2023d). Simple architectures will result in representations with lower dimensions which require less data (Cabannes et al. 2023), while large models tend to overfit with small data (Xie et al. 2023d). The model capacity, the dataset size and the computation resource improve the overall performance according to a power-law relation, unless one factor is bottlenecked by the others (Xie et al. 2023d).

To conclude, the following trends regarding the dependence of network size and data size have been observed: while network size seems to matter more than dataset size, it is most beneficial if both are increased in accordance. "Note that exclusively using more data or larger models may hurt performance; instead, both need to be increased in tandem" (Kolesnikov et al. 2020). The pretext task should be taken into account as it can increase the extraction of data per sample and thereby enable the usage of bigger models. Still, there can be saturation phenomenon with harder downstream tasks and bigger networks as well as pretraining datasets, so there is a limit to upscaling.

8.4 In-domain pretraining

Different domains come with significantly different data characteristics. Medical images e.g. often have a different structure than natural images, including cluttering, no canonical views, specific color distributions, complex textures and fields of views (Anton et al. 2022;

Table 11 Influence of pretraining data and network size on the performance, sorted by learning paradigm (P) (contrastive (C), hybrid (H) and generative (G)), including the parameters (Para) in millions, the downstream datasets (Down data) and the pretraining datasets (1%, 10%, 50% and 100% of ImageNet-1K, ImageNet-21K and ImageNet-21K-P, YFCC-100 M, JFT-300 M, IG-3B). The results with ViT-Tiny originate from Wang et al. (2023d), all contrastive models with ResNet-50 from El-Nouby et al. (2021) and all hybrid models from Lu et al. (2023). The experiments with BEiT or SplitMask and ViT-S stem from El-Nouby et al. (2021), all SimMIM results from Xie et al. (2022a), MAE with Nat18-42K from Singh et al. (2023), MAE with ImNet1K-1.4M from Lu et al. (2023) and the remaining generative results from Zhou et al. (2022b)

Method	P	Model	Para	Down data	KT	ImNet					YFCC	JFT	IG
						14K	140K	700K	1.4M	12/14 M	100 M	300 M	3B
MoCo-v3	C	ViT-Tiny	5.7	ImNet21P-1.2M	FT	76.2	76.5		76.8	**76.9**			
DINO	C	Vit-S	22	iNat19-265K	FT	70.1	73.1		**78.4**				
BYOL	C	ResNet-50	25.6	ImNet1K-1.4M	LP						66.6	**67.6**	
DnC	C	ResNet-50	25.6	ImNet1K-1.4M	LP						67.8	**69.8**	
MoCLR	C	ResNet-50	25.6	ImNet1K-1.4M	LP						65.7	**67.4**	
BYOL	C	ResNet-50	25.6	Places-1.8M	LP						**52.9**	52.4	
DnC	C	ResNet-50	25.6	Places-1.8M	LP						**54.1**	53.3	
MoCLR	C	ResNet-50	25.6	Places-1.8M	LP						53.2	**53.5**	
MAE+DINO	H	ViT-B	86	ImNet1K-1.4M	FT				83.8	**84.0**			
MAE+CMAE	H	ViT-B	86	ImNet1K-1.4M	FT				**83.8**	**83.8**			
MAE+CLIP	H	ViT-B	86	ImNet1K-1.4M	FT				84.4	**84.8**			
MAE	G	ViT-Tiny	5.7	ImNet21P-1.2M	FT	77.9	**78.0**		**78.0**	**78.0**			
BEiT	G	ViT-S	22	iNat19-265K	FT	74.1	74.5		**75.2**				
SplitMask	G	ViT-S	22	iNat19-265K	FT	74.8	**75.4**		**75.4**				
SimMIM	G	SwinV2-S	49	iNat18-42K	FT		76.3	77.3	**77.8**				
MAE	G	ViT-B	86	iNat18-42K	FT				**75.0**				74.7
MAE	G	ViT-B	86	ImNet1K-1.4M	FT				**82.9**	82.7			
BEiT	G	ViT-B	86	ImNet1K-1.4M	FT				83.4	**83.7**			
iBOT	G	ViT-B	86	ImNet1K-1.4M	FT				84.0	**84.4**			
SimMIM	G	SwinV2-B	87	ImNet1K-1.4M	FT		83.2	84.0	85.0	**85.1**			
SimMIM	G	SwinV2-B	87	iNat18-42K	FT		77.1	79.4	79.6	**79.7**			
SimMIM	G	SwinV2-L	195	ImNet1K-1.4M	FT		83.7	85.0	85.8	**85.9**			
SimMIM	G	SwinV2-L	195	iNat18-42K	FT		77.2	81.6	81.8	**81.9**			
MAE	G	ViT-L	307	iNat18-42K	FT				80.2				**80.7**

Table 11 (continued)

Method	P	Model	Para	Down data	KT	ImNet					YFCC	JFT	IG
						14K	140K	700K	1.4M	12/14 M	100 M	300 M	3B
BEiT	G	ViT-L	307	ImNet1K-1.4M	FT				85.2	**86.0**			
iBOT	G	ViT-L	307	ImNet1K-1.4M	FT				85.2	**86.6**			
MAE	G	ViT-H/14	632	iNat18-42K	FT				82.8				**84.0**
SimMIM	G	SwinV2-H	655	ImNet1K-1.4M	FT			85.8	**86.3**	**86.3**			
SimMIM	G	SwinV2-g	1061	ImNet1K-1.4M	FT			85.7	**86.5**	**86.5**			

Table 12 OOD vs. ID pretraining according to the dataset size ratio (see Sect. 8.4 for details and references)

Method	Model	OOD data	ID data	Ratio	Down data	Metric	OOD	ID	Gain	KT
MoCo-v3	ResNet-18	ImNet100	COCO	0.5	CityScapes	mAP	57.4	58.3	1.6	SVM
MoCo-v3	ResNet-18	ImNet100	COCO	0.5	PascalVOC	mAP	56.2	57.6	2.5	SVM
MP-MAE	ConvNeXt-v2	ImNet	MMEarth	1.1	BigEarth20k	Acc	72.9	80.0	9.7	FT
MP-MAE	ConvNeXt-v2	ImNet	MMEarth	1.1	Cashew1k	Acc	77.5	81.4	5.0	FT
MP-MAE	ConvNeXt-v2	ImNet	MMEarth64	1.1	BigEarth20k	Acc	72.9	82.8	13.6	FT
MP-MAE	ConvNeXt-v2	ImNet	MMEarth64	1.1	Cashew1k	Acc	77.5	81.6	5.3	FT
MoCo	ResNet-50	iNat18	ChestX-ray8	3.9	COVID-CT	Acc	79.6	78.6	−1.2	FT
MoCo	ResNet-50	iNat18	ChestX-ray8	3.9	Pneumonia	Acc	94.9	93.3	−1.7	FT
MP-MAE	ConvNeXt-v2	ImNet	MMEarth100k	12.8	BigEarth20k	Acc	72.9	75.5	3.6	FT
MP-MAE	ConvNeXt-v2	ImNet	MMEarth100k	12.8	Cashew1k	Acc	77.5	68.4	−11.7	FT
MoCo-v2	ResNet-50	ImNet	TCGA+TULIP	34.9	BACH	Acc	71.7	85.0	18.6	LP
MoCo-v2	ResNet-50	ImNet	TCGA+TULIP	34.9	CRC	Acc	92.9	93.9	1.2	LP
MoCo-v2	ResNet-50	ImNet	TCGA+TULIP	34.9	MHIST	Acc	79.7	82.3	3.2	LP
MoCo-v2	ResNet-50	ImNet	TCGA+TULIP	34.9	PCam	Acc	83.4	86.5	3.8	LP
MoCo-v2	ResNet-50	ImNet	TCGA	61.0	BACH	Acc	71.7	77.5	8.1	LP
MoCo-v2	ResNet-50	ImNet	TCGA	61.0	CRC	Acc	92.9	93.5	0.7	LP
MoCo-v2	ResNet-50	ImNet	TCGA	61.0	MHIST	Acc	79.7	77.1	−3.3	LP
MoCo-v2	ResNet-50	ImNet	TCGA	61.0	PCam	Acc	83.4	86.8	4.1	LP
SimCLR	ResNet-50	iNat21	ImageNet	76.5	Food101	Acc	65.1	65.7	0.9	LP
SimCLR	ResNet-50	iNat21	ImageNet	76.5	Sport	Acc	94.3	94.3	0.0	LP
MoCo-v3	ViT-B/16	YFCC	ImageNet	76.5	Cars	Acc	25.2	41.9	66.3	LP
SimCLR	ResNet-50	YFCC	ImageNet	76.5	Cars	Acc	37.2	46.6	25.3	LP
MAE	AlexNet	ImNet	fastMRI Knee	128.1	S. Knee MRI	AUC	0.54	0.62	16.4	FT
BYOL	AlexNet	ImNet	fastMRI Knee	128.1	S. Knee MRI	AUC	0.55	0.75	36.4	FT
SimCLR	AlexNet	ImNet	fastMRI Knee	128.1	S. Knee MRI	AUC	0.53	0.80	50.9	FT
MoCo	ResNet-50	ImNet	LUNA	1442.8	COVID-CT	Acc	76.0	79.1	4.1	FT
MoCo	ResNet-50	ImNet	LUNA	1442.8	Pneumonia	Acc	94.1	93.4	−0.7	FT
SimCLR	U-Net	ImNet	DSBCC	2840.7	ACDC CMR	AUC	0.87	0.93	6.0	FT

Table 12 (continued)

Method	Model	OOD data	ID data	Ratio	Down data	Metric	OOD	ID	Gain	KT
MAE	U-Net	ImNet	DSBCC	2840.7	ACDC CMR	AUC	0.84	0.94	11.3	FT
BYOL	U-Net	ImNet	DSBCC	2840.7	ACDC CMR	AUC	0.94	0.94	0.9	FT
BYOL	U-Net	ImNet	DSBCC	2840.7	LV	Dice	0.88	0.93	4.8	FT
MAE	U-Net	ImNet	DSBCC	2840.7	LV	Dice	0.91	0.94	2.5	FT
SimCLR	U-Net	ImNet	DSBCC	2840.7	LV (ACDC)	Dice	0.90	0.93	2.8	FT
BYOL	U-Net	ImNet	DSBCC	2840.7	Myocardium	Dice	0.85	0.88	3.6	FT
SimCLR	U-Net	ImNet	DSBCC	2840.7	Myocardium	Dice	0.87	0.88	1.3	FT
MAE	U-Net	ImNet	DSBCC	2840.7	Myocardium	Dice	0.85	0.89	3.6	FT
SimCLR	U-Net	ImNet	DSBCC	2840.7	RV	Dice	0.88	0.88	0.6	FT
MAE	U-Net	ImNet	DSBCC	2840.7	RV	Dice	0.89	0.88	−0.3	FT
BYOL	U-Net	ImNet	DSBCC	2840.7	RV	Dice	0.88	0.90	1.8	FT

Kang et al. 2023; Pani and Chawla 2025). Consequently, generalizing a pretext task can be challenging if the pretext and downstream data are too different (Dimitrovski et al. 2024; Ericsson et al. 2021, 2022b; Goyal et al. 2019; Van Horn et al. 2021). Instead, SSL excels if pretraining and downstream data are from the same domain (Chen et al. 2020c; Cole et al. 2022; Konstantakos et al. 2025; Sanchez-Fernandez et al. 2024) or if the domain gap is small (Jing and Tian 2020). Still, most SSL algorithms are pretrained on general benchmarking datasets. The large size of such pretraining datasets is supposed to compensate the domain discrepancy (El-Nouby et al. 2021), but there is no general rule regarding the required size for a successful trade-off.

Table 12 compares the performance of self-supervised pretraining using either ID data or out-of-domain (OOD Data) data including the respective pretext tasks method (Method) and model architectures (Models). The performance is evaluated on the downstream data (Down Data) using different metrics (mean Average Precision (mAP), accuracy (Acc), Average Precision$_{50}$ (AP_{50}), AUC, and dice coefficient (Dice)). The column "Ratio" describes the size ratio between the ID and the OOD dataset, "Gain" is the percentage improvement achieved through pretraining with the ID dataset compared to the OOD dataset and "KT" is the knowledge transfer method used. MoCo-v3 results on ResNet-18 originate from Kawashti et al. (2023), who use a support vector machine (SVM) as a linear classifier for feature extraction. MP-MAE results are from Nedungadi et al. (2024), MoCo results stem from Yang et al. (2020a). MoCo-v2 results with TCGA or TCGA+TULIP originate from Kang et al. (2023). SimCLR results with iNat21 as out-of-domain (OOD) dataset have been published by Girish et al. (2022), while Al Kader Hammoud et al. (2025) applied YFCC as OOD data with SimCLR. AlexNet and U-Net results are from Zhang et al. (2024b). On average, ID-pretraining improves the downstream performance by 7.7% according to the data in Table 12. For LP only, the average gain is 10.7%, for FT 6.8% and for SVM 2.1%. For 84.6% of all examples in the table, ID-pretraining is positive or neutral. In all scenarios where the ODD dataset is smaller or not significantly bigger than the ID dataset (ratio 1.1 or below), ID pretraining is beneficial. Still, there seems to be no clear relation between OOD/ID size ratio and performance gain, as the examples with a gain higher than 10% vary from 2840.7 to 1.1 ratio, so ID-pretraining can even be beneficial if the ID-dataset is extremely small compared to the OD-dataset (Zhang et al. 2024b). Additionally, ID dataset can still have very different class similarities, visual distances, etc. There are only few cases where OOD-pretraining is superior, which might depend on the specific domain gap size, which can significantly vary in one domain. Consequently, the specific similarity between the ID pretraining dataset and the downstream dataset might be of more importance than the sole ratio. The benefits of ID-pretraining depend on the specific ID-dataset, e.g. TCGA is disadvantageous for MHIST, but the combination of TCGA and TULIP is beneficial (Kang et al. 2023). Furthermore, ID improvement is not symmetrically; if one pretraining dataset is very beneficial for a downstream dataset of the same domain, the roles cannot always be successfully reversed (Jonske et al. 2025). Design choices involved in the curation of a pretraining dataset should create optimal characteristics for the ID-effect, e.g. a large label granularity, a meaningful label hierarchy and a high alignment (Hong et al. 2023). Additionally, the specific SSL framework and architecture matter. ID-pretraining with TFC5 is e.g. detrimental for OPPD dataset with DenseCL, but advantageous with MoCo-v2 (Ogidi et al. 2023). ViTs in general seem to benefit more from domain-alignment than CNNs (Kang et al. 2023). The ID-effect might also be most relevant for "domain adaptive" pretraining (Lahrichi et al.

2025; Risojević and Stojnić 2022; Zhang et al. 2022b). Upscaling both dataset size and model capacity generates higher performances in ID-training (Filiot et al. 2023). High spatial resolutions are important for ID pretraining with remote sensing images, maybe more important than large datasets and high class diversity (Ghanbarzadeh and Soleimani 2024).

Although there are differences regarding the learning paradigms, ID-pretraining is beneficial in most cases, especially for tasks with a high level of difficulty (Zheng et al. 2021), when the image characteristics of the target domain are very different from those of common pretraining datasets (e.g. in remote sensing (Dimitrovski et al. 2024)) or if the target dataset is rather small (Konstantakos et al. 2025; Sanchez-Fernandez et al. 2024). ID and low-data DINO-pretraining outperforms large-scale supervised pretraining in microscopy imaging, while there is a threshold of dataset size below which SSL is ineffective (50k – 65k images in this case) (Konstantakos et al. 2025). Self-supervised ID pretraining can outperform large-scale pretraining given low data (50k-300k images) (Konstantakos et al. 2025). Similarly, for remote sensing, ID SSL is competitive or even better than OOD SL although the respective OD dataset is 6.5-times bigger (Sanchez-Fernandez et al. 2024). CL rather focuses on high-level structures, probably being able to bridge domain gaps more easily (Zhang et al. 2023c). In contrast, predictive and generative SSL tend to be more sensible regarding overfitting and data bias and therefore might be less useful for high domain gaps (Zhang et al. 2023c). Due to the similarity of ImageNet and common downstream datasets (e.g. Cars, Flowers, Pets or Food), the high performance could partly be explained with the ID-effect (Al Kader Hammoud et al. 2025).

However, it is essential to understand the causes of the ID-pretraining benefits. One aspect is the pretext data bias (Purushwalkam and Gupta 2020), which can be detrimental given significantly different downstream data characteristics. If the information extracted during the pretext task is not beneficial for the target task, negative transfer can occur, as poorly performing, rare or erroneously correlated features can be transferred, which hinders successful generalization (Yang et al. 2023b). As domain properties are reflected by Batch Normalization (BN) statistics, a large domain gap leads to the development of an unfavorable mixed data distribution within the BN layers during FT (Zhang et al. 2023c; Zhou et al. 2022a). In contrast, ID pretraining results in a higher feature reuse rate and is also more beneficial for complex downstream tasks, smaller downstream datasets or in general when LP is performed (Jonske et al. 2025). ID-models also seem to have a "lower attentive diffusion", they only encode key areas of the ID-data and are characterized by "focused feature extraction" (Anton et al. 2022). ID pretraining in low-data regimes for medical images is probably beneficial as the features of large-scale RGB datasets are not helpful for these tasks and SSL is less affected by class imbalance issues (Konstantakos et al. 2025).

In summary, SSL pretraining is most beneficial when source and target domain are similar, while small downstream datasets tend to benefit from SSL pretraining even if the domain gap is wider (Al Kader Hammoud et al. 2025; Matsoukas et al. 2022). In cases where image characteristics vary significantly within one domain, ID pretraining may be less helpful than training with general benchmarking datasets (Jonske et al. 2025). The success of ID-pretraining raises the question if the design of suitable pretraining datasets is more important than pretext task optimization (Gui et al. 2024).

8.5 Self-pretraining

Self-pretraining is a sub-type of ID-pretraining, where pretraining is executed with the downstream dataset to avoid domain discrepancies (Jiang and Veeraraghavan 2024; Krishna et al. 2023; Zhou et al. 2023a). Due to the effectiveness of ID-pretraining, the question arises whether the more extreme self-pretraining is beneficial, as there is no benefit in using a different pretraining dataset if all the necessary visual features can already be extracted from the downstream data (Mensink et al. 2022). Self-pretraining is e.g. common in medical image analysis (Jiang and Veeraraghavan 2024). Table 13 compares self-pretraining with standard pretraining taking the size ratio of both datasets into consideration. On average, self-pretraining improves the downstream performance by 14.5%. According to the results of Table 13, self-pretraining is always significantly better if pretraining and self-pretraining dataset have the same size: Cole et al. (2022) pretrain SimCLR on iNat21, ImageNet, Places365 and GLC20 using only 1 M of the full datasets to generate equal sizes for both pretraining and self-pretraining. They evaluate the downstream performance using LP and the full datasets (Cole et al. 2022). The generated improvements due to self-pretraining are very high (up to 163.6%) due to the bad pretraining performance. For a size ratio of up to 16.9, self-pretraining is beneficial in most cases and slightly negative in few. For ratios of 21.0 or higher, self-pretraining is mostly detrimental. Still, when using Derm dataset for downstream classification and ImageNet for pretraining with a ratio of 83.5, self-pretraining is better, probably due to the high domain gap (Azizi et al. 2021). The exact gain is hard to forecast, as the bigger CheXpert (Irvin et al. 2019) profits way more from ImageNet pretraining than Derm (Azizi et al. 2021). Still, there seems to be a general trade-off between domain gap and data scarcity.

Jiang and Veeraraghavan (2024) compare standard pretraining with self-pretraining on both Swin and ViT regarding lung tumors segmentation with SMIT and PRCL-v2. Pretraining is done with a large dataset (10,412 3D CT scans) with subsequent FT on 350 scans, while self-pretraining only used these 350 scans (Jiang and Veeraraghavan 2024). To compensate for this difference, self-pretraining is four times longer (Jiang and Veeraraghavan 2024). Self-pretraining performs worse, but both the huge difference in dataset size and the fact that the pretraining dataset and the self-pretraining dataset are from the same domain have to be taken into consideration (Jiang and Veeraraghavan 2024). Similar to ID-pretraining, the effectiveness of self-pretraining seems to depend on whether the higher similarity can compensate for the smaller dataset size, as self-pretraining with small datasets can cause overfitting (El-Nouby et al. 2024). However, it is difficult to determine the specific dataset ratio for the trade-off. Self-pretraining with Flowers dataset is detrimental, which could be attributed to the low diversity of the dataset with high inter-class similarity used for fine-grained classification (Yangyang and Xiang 2019), which probably makes it relatively unsuitable for pretraining. Compared to ImageNet pretraining, self-pretraining generates on average 4.9% improvement, while compared to all other pretraining datasets the improvement is 20.1%. This demonstrates the general suitability of ImageNet for pretraining. In general, self-pretraining seems to decrease performance most if both the downstream dataset and the domain gap are small. Still, it could be a good option given sufficiently large downstream data and a big domain gap. Consequently, self-pretraining is particularly useful if only an OOD dataset is available as alternative. If a larger ID dataset is available, this should probably be preferred. This conclusion is consistent with the findings

Table 13 Improvements (Gain) due to self-pretraining (Self-pretrain) compared to standard pretraining (Pretrain) sorted by the ratio of the respective dataset sizes (Ratio), including pretext tasks (Method), model architecture (Architecture) and KT method (KT). The self-pretraining dataset is always used for downstream training. Evaluation metrics are mean average precision (mAP), accuracy (Acc), AUC, mask average precision (AP^m), box average precision (AP^b), and mean Intersection-over-Union (mIoU). SimCLR results with LP are from Cole et al. (2022) who use only 1 M of the respective dataset for both types of pretraining, but the full datasets for the downstream task. SimCLR results with FT originate from Azizi et al. (2021), MAE results from Hu et al. (2022) and MoCo results from Yang et al. (2020a). Results with SMIT and PRCL-v2 are from Jiang and Veeraraghavan (2024), result with AIM from El-Nouby et al. (2024)

Method	Architecture	KT	Pretrain data	Self-pretrain data	Ratio	Metric	Pretrain	Self-pretrain	Gain
SimCLR	ResNet-50	LP	iNat21 (1 M)	GLC20 (1 M/Full)	1.0	mAP	70.7	76.9	13.1
SimCLR	ResNet-50	LP	iNat21 (1 M)	Places365 (1 M/Full)	1.0	Acc	41.6	50.0	17.1
SimCLR	ResNet-50	LP	Places365 (1 M)	GLC20 (1 M/Full)	1.0	mAP	18.7	76.9	126.5
SimCLR	ResNet-50	LP	Places365 (1 M)	iNat21 (1 M/Full)	1.0	Acc	29.2	49.3	25.7
SimCLR	ResNet-50	LP	GLC20 (1 M)	iNat21 (1 M/Full)	1.0	Acc	18.7	49.9	163.6
SimCLR	ResNet-50	LP	GLC20 (1 M)	Places (1 M/Full)	1.0	Acc	32.9	50.1	52.3
SimCLR	ResNet-50	LP	ImageNet (1 M)	iNat21 (1 M/Full)	1.0	Acc	37.3	49.3	32.2
SimCLR	ResNet-50	LP	ImageNet (1 M)	Places365 (1 M/Full)	1.0	Acc	48.6	50.1	3.1
SimCLR	ResNet-50	LP	ImageNet (1 M)	GLC20 (1 M/Full)	1.0	mAP	71.6	76.9	7.4
MAE	ViT-B	FT	ImageNet	COCO	5.3	AP^m	44.6	44.9	0.7
SimCLR	ResNet-50	FT	ImageNet	CheXpert	5.7	AUC	0.763	0.765	0.2
SimCLR	ResNet-50(4x)	FT	ImageNet	CheXpert	5.7	AUC	0.768	0.767	−0.2
SimCLR	ResNet-152(2x)	FT	ImageNet	CheXpert	5.7	AUC	0.767	0.768	0.2
MAE	ViT-L	FT	Places365	COCO	7.5	AP^m	46.6	47.1	1.1
MAE	ViT-B	FT	Places365	COCO	7.5	AP^b	49.2	50.5	2.6
MAE	ViT-L	FT	Open Images	COCO	14.7	AP^m	47.3	47.1	−0.4
MoCo	ResNet-50	FT	SUN-397	Caltech-256	16.9	Acc	53.9	56.6	5.0
MoCo	ResNet-50	FT	SUN-397	Flowers-102	21.0	Acc	78.1	57.6	−26.3
MoCo	ResNet-50	FT	SUN-397	Pneumonia	22.3	Acc	93.8	93.4	−0.3
SMIT	ViT-B	FT	3D CT/LC	TN/LRad	29.7	Acc	66.0	64.0	−3.0
SMIT	Swin-S	FT	3D CT/LC	TN/LRad	29.7	Acc	69.0	63.0	−8.7
PRCL-v2	nsUnet	FT	3D CT/LC	TN/LRad	29.7	Acc	46.0	45.0	−2.2
MAE	ViT-B	FT	ImageNet	ADE20k	63.4	mIoU	47.5	48.2	1.5
SimCLR	ResNet-50	FT	ImageNet	Derm	83.5	Acc	62.6	63.7	1.7

Table 13 (continued)

Method	Architecture	KT	Pretrain data	Self-pretrain data	Ratio	Metric	Pretrain	Self-pretrain	Gain
SimCLR	ResNet-50(4x)	FT	ImageNet	Derm	83.5	Acc	64.6	66.9	3.6
SimCLR	ResNet-152(2x)	FT	ImageNet	Derm	83.5	Acc	66.4	66.4	0.1
MoCo	ResNet-50	FT	SUN-397	COVID-CT	374.0	Acc	79.4	61.7	−22.2
AIM	AIM-0.6B	LP	DFN-2B	ImageNet	1506.4	Acc	74.5	73.5	−1.3

for self-pretraining and NLP (Krishna et al. 2023). Clear threshold values cannot be determined as these are probably also dependent on other design choices and data characteristics.

9 Predictive learning

The choice of a suitable pretext task and the respective hyperparameters is essential to enable the network to learn semantically meaningful representations with predictive SSL and requires both domain adaption and knowledge (Jing and Tian 2020). Given different predictive approaches, the pretext task type seems to be more important than the network size: Comparing Jigsaw (ResNet-50 2x, 94 M), Rel. Position (ResNet-50 2x, 94 M) and Rotation (RevNet-50 4x, 86 M) for ImageNet classification, the smallest model has the highest performance 55.4%, followed by Rel. Position (51.4%) and Jigsaw (44.6%) (Wang et al. 2022a). However, the group of predictive SSLs is very heterogeneous, making generalizations difficult. The complexity and consequently the difficulty of the pretext tasks are an important factors for downstream performance (see 6.1). Predictive SSL is based on pseudo-labels generating pseudo-categories, so both difficulty and performance are highly influenced by how easily distinguishable these are (see Sect. 9.1). While CL and MIM include projectors or heads being discarded after pretraining as intermediate representations are more useful for downstream learning (Alkin et al. 2025; Ericsson et al. 2022b), for predictive SSL the question which layer to extract the features from is very relevant (see Sect. 9.2). Many predictive SSL models also combine multiple pretext task (see Table 9).

9.1 Pretext difficulty and distinguishable pseudo-categories

Most predictive pretext tasks are are influenced by the difficulty and the number of distinguishable pseudo-categories of the pretext task. Jigsaw includes e.g. a hyperparameter determining the number of patch permutations; higher values increase the difficulty and thereby the performance (Goyal et al. 2019; Noroozi and Favaro 2016). Still, the maximum value is restricted due to the number of learnable parameters in the output layer growing linearly with the number of patch permutations (Misra and Maaten 2020). Rel. Position has been developed to teach the model the spatial context of the objects (Doersch et al. 2015). Locating objects in different patches poses the danger of failing to learn the relationship between patches (Wang et al. 2023a). Furthermore, instead of semantically meaningful features, trivial ones can be learned and global information is excluded (Wang et al. 2023a). Shortcut learning via texture continuity or boundary patterns has to be avoided by increasing the pretext difficulty e.g. by additional gaps, color channel shifting, random patch displacement, or randomly jittered patches (Shurrab and Duwairi 2022; Wang et al. 2023a). Colorization requires pixel-level and semantic scene understanding, resulting in representations which are useful for object classification, detection, and segmentation (Jin et al. 2020; Larsson et al. 2016; Xiao et al. 2022; Zhang et al. 2016). Predicting color histograms is superior to a single color prediction, as many objects have more than one possible color (Larsson et al. 2016; Zhang et al. 2016). Colorization profits from complex networks, while complex architectures benefit from more complex colorization tasks (Larsson et al. 2016).

Although being a rather simple pretext task, Rotation helps the model to deal with different views, captures high-level features and requires semantic understanding of objects as

well as their locations and backgrounds (Jing et al. 2018). Furthermore, it does not increase computation complexity excessively and avoids learning trivial solutions, as it generates no artifacts (Jing et al. 2018). Rotation is only beneficial if learning rotation-invariant features contributes to the downstream task and if the dataset has a canonical orientation. The optimal number of rotations for 2D images is consistent across tasks: Four rotational modes (0°, 90°, 180°, 270°) appear to be a good trade-off between too few classes and indistinguishable orientations (Gidaris et al. 2018; Ji et al. 2022; Jing et al. 2018; Yamaguchi et al. 2021; Yi et al. 2022). For Rotation with 3D point clouds, Poursaeed et al. (2020) use Orientation Estimation and consider each point cloud as a regular polyhedron with several vertices. They achieve the best results using six rotation angles, while each can be understood as a direction on the three axes (Poursaeed et al. 2020). When applying more rotation angles, the accuracy decreases significantly, e.g. for 18 rotations angles it drops from 99.8% to 89.0% (Poursaeed et al. 2020). These results show the importance of clearly distinguishable pseudo-categories for predictive learning.

9.2 Feature extraction layer

Choosing the best feature extraction layer for the downstream task is particularly crucial for predictive SSL (Ericsson et al. 2022b; Goyal et al. 2019), as the models mostly do not use a projector or head which is removed after pretraining. While the first layers encode more simple representations and later layers increasingly complex and abstract patterns, the best option is task- and data-specific and could profit from a combination of features from various layers (Ericsson et al. 2022b; Gidaris et al. 2018; Jing and Tian 2020). The representation quality can decrease in deeper layers if the model loses general semantic features and becomes too specialized to the pretext task (Gidaris et al. 2018; Kolesnikov et al. 2019; Noroozi and Favaro 2016). Predictive SSL tends to be as good as SL in the first layers, while rather struggling with deeper layers (Mundhenk et al. 2018). For AlexNet on ImageNet, the best accuracies are achieved in layer three or four and the performance decreases again in the last layer (Kim et al. 2018). For the medical NCC dataset, U-Net based architectures and different predictive frameworks (Rotation, RPL, Rubik's Cube), features from the first three layers are most useful for the downstream task (Zhang et al. 2023c). Similar observations regarding the superiority of intermediate layers are provided by Caron et al. (2018); Zhang et al. (2017); Zhuang et al. (2019b) and are in accordance with similar trends for contrastive and generative SSL.

10 Contrastive learning

The most important elements of CL regarding design choices are data augmentation, encoder, and projector (see Fig. 9). CL approaches with negative samples require additional sampling. Together, negative and positive samples determine which features are discriminated and dominate learning (Robinson et al. 2021b).

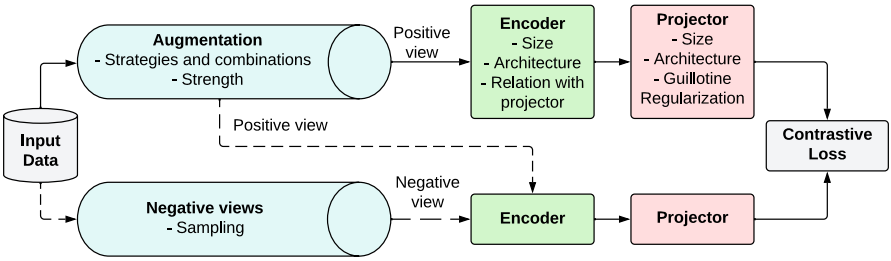

Fig. 9 Important design choices for CL include the selection of augmentation strategies, their combination and strength. CL methods with negative views require negative sampling strategies. The size and architecture of encoder and projector, the encoder-projector relation and Guillotine regularization should also be considered

10.1 Data augmentation strategy

CL is very sensitive regarding data augmentation (Albelwi 2022; Giakoumoglou et al. 2025; Grill et al. 2020), which is an integral part of the pretext task as it creates positive (and negative) views, which allow learning view-invariant representations (Hu et al. 2024b; Zhang et al. 2024f). MIM are more sample-efficient and consequently require less augmentations (El-Nouby et al. 2024; Moutakanni et al. 2024). Data augmentations implicitly encode weak supervision (Bendidi et al. 2024) and shift the focus from instance classes to a direct comparison of image features with the applied transformations defining the learned invariances (Caron et al. 2021a). Augmentations are crucial for sparsity, as the network learns feature representations focusing on spurious dense noise otherwise (Wen and Li 2021). They enable "feature decoupling", so models concentrate on the features associated with the sparse signals instead of the ones related to noise (Jia and Chan 2022; Wen and Li 2021). Augmentations exploit domain-specific inductive biases such as symmetries or equivariances (Chen et al. 2021b). Table 14 lists important contrastive models and the augmentations they apply, showcasing that most use similar combinations.

According to Morningstar et al. (2024), enhanced augmentation strategies can generate improvements of 2-4%, while algorithmic enhancements (such as novel loss functions or incorporating prediction networks) generate improvements below 1%. According to their work, optimizing data augmentation techniques is a more important driver of SSL success than optimizing pretraining, as the main performance gain when comparing older frameworks (e.g. SimCLR) and newer ones (e.g. DINO) can be attributed to augmentations (5%), followed by model size (2.3%), momentum encoder (1.8%), and predictor (0.3%) (Morningstar et al. 2024). Zhang et al. (2021e) show that optimizing data-augmentation-related hyperparameters can have a greater effect than optimizing classical training hyperparameters such as learning rate, optimizer or weight decay. Wagner et al. (2022) compare the impact of diverse hyperparameters (learning rate, warmup epochs, warmup multiplier, optimizer, weight decay start and end) and augmentation strategies (e.g. color, quality and geometric transformations) on the model performance using SimSiam, ResNet-18 as well as the datasets CIFAR-10, CIFAR-100 and DermaMNIST. While hyperparameter optimization only results in marginal improvements, optimizing the data augmentation strategies increases performance by at least 1% to 2.3% (Wagner et al. 2022). In turn, optimizing the hyperpa-

Table 14 Augmentation combinations of popular CL models including the data augmentation strategies Random Cropping (Crop), Random Color Jitter (Jitter), Random Color Distortion or Grayscale (RCD/G), Random Horizontal Flip (Flip), Random Gaussian Blur (Blur), Random Solarization (Solar), Fast AutoAugment (AA), Color Normalization (CN), and Uniform Noise (UN). Cropping strategies can include multi-cropping (Multi) or multi-local cropping. MAE-CT is a hybrid approach

Method	Crop	Jitter	RCD/G	Flip	Blur	Solar	AA	CN	UN
DeepCluster	x			x					
AMDIM	x	x	G	x			x		
MoCo	x	x	G	x					
BYOL	x	x	RCD	x	x	x			
MoCo-v2	x	x	RCD	x	x				
SimCLR	x	x	RCD	x	x				
SimCLR-v2	x	x	RCD	x	x				
SimSiam	x	x	G	x	x				
BT	x	x	G	x	x	x			
DINO	Multi	x	G	x	x	x			
MoCo v3	x	x	G	x	x	x			
NNCLR	x	x	RCD	x	x	x			
SwAv	Multi	x	RCD	x	x				
Mugs	Multi	x	G	x	x	x	x		
SimCL									x
VICReg	x	x	G	x	x	x		x	
AUC-CL	Multi	x	G	x					
Dino-MC	Multi-local	x	G		x				
MAE-CT$_{min}$	x			x					
MAE-CT$_{aug}$	x	x	RCD	x	x	x			

rameters of the augmentations can achieve significant performance improvements (Bendidi et al. 2024; Chen et al. 2020b).

But what characterizes optimal augmentation strategies? Each positive view should be significantly different from all other images (Pal et al. 2019). The distribution of augmented images should have a low inter-class but a high intra-class overlap (Abnar et al. 2021). Optimal views are task-dependent and influence the learned semantics (Balestriero et al. 2023; Tian et al. 2020b). The downstream task mainly benefits from polar (in-)variances, which is why data augmentation should be precisely adapted to the specific downstream feature (in-) variances required (Ericsson et al. 2022a). Different methods of CL learn different invariances as they apply different transformations, which in turn leads to features being invariant to that augmentation (Balestriero et al. 2023; Ericsson et al. 2022a). These invariances generalize to real-world transformations and strongly affect downstream performance (Ericsson et al. 2022a). The views should only "share the minimal information" required for the downstream task and the MI between views should exclude nuisance variables (Tian et al. 2020b). This requires massive changes in the visual appearance of images without manipulation of the semantic meaning (Albelwi 2022). Transformation should preserve all required class information and therefore can be understood as a form of weak supervision (Bendidi et al. 2023). Consequently, the appropriate data augmentation strategies are domain-dependent and require domain expertise (Bendidi et al. 2023). As a result, recipes developed for standard RGB-images benchmarking datasets (mostly ImageNet) might be less suitable for other datasets (Bordes et al. 2023b; Scheibenreif et al. 2022; Xiao et al. 2021). Furthermore, hand-crafted data augmentations are especially relevant given bigger model sizes, as more augmentation can mitigate the risk of overfitting (Moutakanni et al. 2024). Increasing the data augmentations is therefore an alternative to upscaling the pretraining dataset size with the network size (Moutakanni et al. 2024). According to Liang et al. (2023), applying semantic augmentations to each modality helps to preserve task-relevant information in multimodal settings. Multimodal augmentations are superior to unimodal ones (Liang et al. 2023). However, augmentations removing information which is shared across modalities should be avoided (Liang et al. 2023). According to Dufumier et al. (2025), strong augmentations are important to learn multimodal interactions. In the following sections, the most important data augmentation strategies and how to combine them are discussed.

10.1.1 Color augmentations

The most important color augmentations for CL are distortion, grayscale, and jitter. Chen et al. (2020b) define color distortion as a combination of color dropping, brightness, contrast, saturation, and hue. While having a huge influence on natural images (Chen et al. 2020b; Wang et al. 2022a), the benefits seem to be similar to simple grayscale (Chen et al. 2020a). Grayscale seems to be essential for both SimCLR and BT (Bordes et al. 2023b). Both are mostly used in combination with color jitter, which randomly adjusts hue, contrast, brightness and saturation (Balestriero et al. 2023) and consequently produce samples which are different from natural color changes (Purushwalkam and Gupta 2020; Zini et al. 2023). Both color distortion and jitter result in a loss of hue information while preserving object positions (Von Kügelgen et al. 2021). Self-pretraining results on CIFAR-100 with SimSiam drop by 18% in accuracy when removing all color augmentations, as less color invariance can be extracted by the encoder (Zini et al. 2023). With SimCLR, color modification and

cropping are the most effective strategies among crop, cutout, color, sobel, noise, blur and rotate regarding ImageNet Top-1 accuracy Chen et al. (2020b). In contrast, CIFAR-pre-training combined with Flowers-102 downstream classification profits from removing color augmentations (12.75% accuracy increase) as color augmentations can prevent CL models from learning color information, which is problematic for downstream tasks requiring this knowledge (Cole et al. 2022; Ericsson et al. 2021; Zini et al. 2023), demonstrating the data-set- and downstream dependency of suitable augmentation choices.

10.1.2 Random resized cropping

Random resized cropping (Crop) is one of the most common augmentation techniques, mostly center crops with a resized final resolution of 224x224 are applied (Grill et al. 2020). SimCLR, BYOL and SimSiam are very sensitive regarding the removal of Crop (Berg et al. 2022; Dangovski et al. 2021). Due to Crop, position information is discarded (Von Kügelgen et al. 2021) which is only useful if the image patches have a high amount of shared information (Tamkin et al. 2021). Presumably, Crop only leads to beneficial representation learning on object-centric datasets (Purushwalkam and Gupta 2020). The crop size should be in accordance with the required scale of details, fine-grained images require smaller crop sizes (Cabannes et al. 2023). Crop shifts the focus to details within the crop size (Cabannes et al. 2023). Eliminating long-range interactions between image parts can lead to erroneous features but causes a forced locality which is often beneficial for generalization (Cabannes et al. 2023). SwAV introduces Multi-Crop, which combines small crops (96 x 96) with two big ones (224 $\times$ 224) and compares the big crops with all others (Caron et al. 2021a), creating a "local-to-global" augmentation (Zhou et al. 2022c). This multi-view strategy can improve both robustness and generalization as one sample can be observed from different perspectives, improving the quality of the extracted features (Hu et al. 2024b). Models can consequently handle deformation, occlusion, and illumination better (Hu et al. 2024b). While SL does not benefit from Multi-Crop, it does improve most CL frameworks (Caron et al. 2021a) and is beneficial for feature quality (Caron et al. 2021b). According to Balestriero et al. (2023), varying the number and size of the crops can increase the performance between 0.3 and 4 percentage points. Overlaps of local and global crops can furthermore be beneficial (Hu et al. 2025). Morningstar et al. (2024) report an average improvement of 2.5% top-1 LP accuracy due to Multi-Crop with SimCLR, BYOL, MoCo v2, SwAV, DINO, and MoCo-v3 on ViT-B, while DINO (+9.6%) and SwAV (+3.6%) profit most.

10.1.3 Advanced augmentation techniques

Several authors aim to use alternatives to handcrafted data augmentation techniques, such as reconstruction-based objectives, which avoids having to define specific invariances (Assran et al. 2023; Baevski et al. 2023; He et al. 2022), style-based augmentations, which try to store the style information contained in the augmentations (e.g. color) to improve downstream tasks that require such information (Dangovski et al. 2021; Gidaris et al. 2018; Scherr et al. 2022; Xiao et al. 2021), equivariance encoding (Marchetti et al. 2023; Park et al. 2022), splitting representations as invariant and equivariant (Garrido et al. 2023b; Ibrahim et al. 2022) (Balestriero et al. 2023), learning hyper-representations of the network (Schürholt et al. 2022), AutoAugment (Cubuk et al. 2018) or Fast AutoAugment (Lim et al. 2019).

Zhang et al. (2023f) propose changing the applied augmentation over the course of the training time, which improves classification of a subset of ImageNet with 100 classes by 1.1% on MoCo-v2 using LP. Zhang et al. (2024f) augment in the feature space for domain-agnostic results. Training a generative model to produce useful views for a given objective might be a strategy for more general CL without domain expertise or tailored pretraining (Tamkin et al. 2021). As the standard transformations applied are rather simple and not representative for natural image variations, which might constrain both diversity and quality of the learned representations, generative augmentation could be one solution to guard the semantics of the source image (Ayromlou et al. 2025). Still, they work best in combination with standard augmentations (Ayromlou et al. 2025). Furthermore, there are several approaches to generate positive samples without augmentation (Hu et al. 2024b): Quan et al. (2023) exploit the semantic context and use background information as negative samples, Dwibedi et al. (2021) select the most similar positive sample within a support set and Park et al. (2020) divide images into patches which are used as positive and negative samples.

10.1.4 Combining data augmentation strategies

Combining multiple data augmentations is the best strategy for effective representation learning: it improves results and generates more general features, prevents low-level shortcuts and speeds up convergence (Cabannes et al. 2023; Chen et al. 2020b; Ohri and Kumar 2021). Still, more original samples are even better than more views (Bardes et al. 2022; Cabannes et al. 2023). The combination of Crop, Color Distortion, Color Jitter, Horizontal Flip, and Gaussian Blur is fundamental for most instance-wise CL frameworks (see Table 14, Bachman et al. (2019) and Bai et al. (2022)). Crop and then Color Distortion leads to 56.3% top-1 LP-accuracy on ImageNet with SimCLR, while sole Crop results in 33.1% and Sole Color Distortion in 18.8% (Chen et al. 2020b). According to Bordes et al. (2023b), the combination of Crop and Grayscale already yields very good performance for both SimCLR and BT with ImageNet (63.5% top-1 accuracy with BT). With BYOL augmentations, Self-Classifier achieves 68.1% ImageNet Top-1 LP-accuracy after 100 epochs with Color Jittering, Gaussian Blur and Solarization, but they reach 72.4% if Multi-Crop and Nearest Neighbour Augmentation (Dwibedi et al. 2021) are added (Amrani et al. 2022). Morningstar et al. (2024) evaluate the impact of more diverse data augmentations on ImageNet top-1 LP-accuracy for SimCLR, BYOL, MoCo v2, SwAV, DINO, and MoCo v3 on ViT-B. More diverse augmentations always seem to be beneficial, the relative increase can be up to 23% compared to sole Crop or 5% compared to the SimCLR augmentations (Morningstar et al. 2024). BYOL augmentations increase performance on average by 2.3% compared to SimCLR augmentations, although the only difference is the addition of Solarization and minor hyperparameter adjustments, which emphasizes the importance of augmentation optimization (Morningstar et al. 2024).

While ImageNet, Cifar10, iNaturalist18 and Places205 clearly profit from applying Solarization with SimCLR; CLEVR-Dist and Eurosat do not (Bordes et al. 2023b). Zhang et al. (2023c) analyze the benefits of additional data augmentation (Flip, Rotation, and Random Scaling) on medical data with SimCLR and BYOL. They conclude that strong additional augmentations degrade the performance, especially for SimCLR, while both models benefit from additional Rotation (Zhang et al. 2023c). Rotation decreases the downstream capabilities to classify rotation angles (Xiao et al. 2021), which might help with rotation-

invariant data (Jaiswal et al. 2020). Removing Color Distortion decreases top-1 accuracy on ImageNet by 9.1 percentage points with BYOL, but 22.2 with SimCLR (Grill et al. 2020). SwAV and BT both are sensitive to Crop: When training on BigEarthNet without Crop, they collapse, while MoCo-v2 and SimSiam only suffer from a performance drop (Wang et al. 2022a). Consequently, CL frameworks differ significantly regarding their sensitivity towards the removal of specific data augmentation techniques.

10.1.5 Strength of data augmentations

Stronger augmentations can result in more structured representations, reducing the search space and preventing collapses (Cabannes et al. 2023). Harder augmentation methods (e.g. AutoAugment, Fast AutoAugment or Jigsaw) are especially effective for CL with negative samples (Bachman et al. 2019; Bai et al. 2022; Tian et al. 2020b; Zhou et al. 2022c). The appropriate complexity level of finding similarities in the augmented views has to be high enough to compensate for the labels, but the image should still be representative of the original distribution (Albelwi 2022) to avoid performance degradation and collapsed solutions (Bai et al. 2022). Strong data augmentations help to learn effective features (Zhang et al. 2023f) and improve downstream classification accuracy by decreasing the MI between views (Tian et al. 2020b). The augmentation strength determines the degree of invariance of the representations and how strongly each view and its nearest neighbor should be projected to opposite directions (Cosentino et al. 2022a). The strength determines if the projector rather than the encoder learns to be invariant to augmentations (Cosentino et al. 2022a). Stronger and more diverse augmentations result in a higher tendency of the projector for dimensional collapse, which makes the encoder more suitable for the downstream task (Cosentino et al. 2022a). Cosentino et al. (2022a) propose an upper bound to InfoNCE loss which can be split into an invariance loss, a repulsion loss and a logarithmic term given a linear projection. Given light augmentations, the repulsion term dominates the loss function while stronger augmentations balance invariance and repulsion loss (Cosentino et al. 2022a). While too hard data augmentation can be counterproductive in supervised settings as class boundaries can be transgressed, hard positive samples in SSL improve instance discrimination (Appalaraju et al. 2020). Bachman et al. (2019) compare the effects of augmentation strength, NCE loss regularization and multiscale feature learning: They use STL10 and ImageNet with AMDIM and conclude that switching from basic to strong augmentations has the strongest effect. Still, strong augmentations come with disadvantages: due to the massive Crop in MoCo and PIRL, the learned representations are inferior at capturing viewpoint and instance invariance, which can hinder occlusion-invariant learning and object discrimination in non-object-centric dataset (Appalaraju et al. 2020; Purushwalkam and Gupta 2020). The network should not merely rely on the augmentation patterns instead of the underlying representation (Zhang et al. 2023f). Strong augmentations when training on the downstream task can degrade segmentation performance given medical images (Zhang et al. 2023c). Tian et al. (2020b) analyze the influence of the strength of Crop and Jitter on the ImageNet downstream accuracy: both methods show a reverse-U shape, indicating that medium values provide the best results. The strength of the augmentation must therefore be well tuned to enable successful CL. According to Bendidi et al. (2024), both the augmentation probability and intensity impact the accuracies of classes differently, as different classes are characterized by a different level of granularity and uniformity and different transfor-

mations could corrupt class-defining information. In conclusion, stronger augmentations improve probe generalization and representation as they improve the invariance of the SSL encoder, thereby reducing the amount of task-irrelevant information that probes could otherwise overfit on (Dubois et al. 2024; Mitrovic et al. 2020; Tian et al. 2020b; Tsai et al. 2021). They also increase the representation usability by decreasing the required capacity of the probes (Dubois et al. 2024, 2022).

10.2 Negative samples

Negative samples should help the model to differentiate between categories, to learn more discriminative features and to generate an improved understanding of the data complexity and diversity (Hu et al. 2024b). Sampling the negative views has to be in accordance with the downstream objective (Hu et al. 2024b) and the dataset to maximize benefits and avoid misleading effects. Negative examples that are very similar to semantically positive ones are referred to as false negative examples (Huynh et al. 2022). They result in relevant information being discarded, slow convergence and decreased performance (Hu et al. 2024b; Huynh et al. 2022). Several strategies have been developed to detect and discard false negatives (e.g. Auh et al. (2024); Balmaseda et al. (2025); Huynh et al. (2022)). While data augmentation mostly determines the characteristics of the positive views, negative sampling is essential for those frameworks that include negative samples (Liu et al. 2023e; Wagner et al. 2022). Harder samples in general increase the pretext task complexity (Dong et al. 2024). The number of available negative samples can have a significant influence, which makes the batch size an important factor (Chen et al. 2020b; He et al. 2020; Hu et al. 2024b; Misra and Maaten 2020; Wu et al. 2018). The presence of hard negative samples influences the balance between the attraction of positive and the repulsion of negative samples (Manna et al. 2024). Still, an excess of negative samples can result in negative pairs being semantically similar and cause performance degradation (Arora et al. 2019; Li et al. 2021e). If their proximity in the embedding space is high, class collision can occur and impede the learning of global semantic structures, as the deviations within one class could become more important than the class differences (Arora et al. 2019; Li et al. 2021e). Augmenting negative samples (Bordes et al. 2023b) and hard sampling of negative views (Robinson et al. 2021a; Zhang et al. 2022a) are possibilities to solve class collision problems. Negative augmentation strategies can also be optimized for specific data characteristics (see e.g. Hu et al. (2024a) for asymmetric augmentation for fine-grained objects). InfoNCE with cosine similarity dependent temperature scaling can optimize sample distribution in the feature space (Manna et al. 2024).

10.3 Projector

Many CL frameworks use a projector (or projection head) on top of the encoder (also backbone or feature extractor) (Cosentino et al. 2022b; Hu et al. 2024b; Ozbulak et al. 2023; Wang et al. 2020). The projector network has a shallow architecture (see Sect. 10.3.2), is destined for pretraining loss calculation and discarded after pretraining (Balestriero et al. 2023; Gupta et al. 2022; Ma et al. 2023). The introduction of projectors stems from the observation that the last layer of a network is not always the ideal to extract knowledge for the actual task (Chen et al. 2020c). Lower encoder layers also still include subclass-

level features (Xue et al. 2024). Consequently, downstream learning is based on the encoder representations which often transfer and generalize better due to the reduced pretraining bias (Cosentino et al. 2022a; Mialon et al. 2024; Przewikeźlikowski et al. 2024; Shwartz Ziv and LeCun 2024; Xue et al. 2024). Projectors increase downstream accuracy significantly (Bordes et al. 2023a), e.g. with VICReg and ImageNet, accuracy is improved by 20 percentage points when adding a projector during pretraining (Bardes et al. 2022). Downstream training without the projector can diminish the accidental retention of details specific to images (déjà vu memorization) as earlier layers have a more comparable inference accuracy of target and reference model (Meehan et al. 2023).

Projectors buffer between training loss and encoder (Ohri and Kumar 2021), thereby mitigating optimization problems and absorbing "any bias related to an ill optimization" (Bordes et al. 2023a). Therefore, projectors prevent training collapse (Ma et al. 2023), overfitting to the pretext task and foster pairwise independence among features (Balestriero et al. 2023; Mialon et al. 2024). They separate "the information-rich features useful for downstream tasks (i.e. color, rotation, or shape of objects) from the more discriminative features that are more useful for the contrastive loss" (Appalaraju et al. 2020). Projectors choose the best subspace to apply the contrastive loss to and can mitigate some of the shortcomings of loss optimization (Gupta et al. 2022). While the projector can in general avoid misalignment between pretext and downstream task, better matching of the two tasks can also render the projector unnecessary (Balestriero et al. 2023; Bordes et al. 2023a; Sariyildiz et al. 2023). According to Mialon et al. (2024), training the projector is not useful if no invariance is required, as projectors with random weights are already able to achieve pairwise independent features, which is favorable to disentangle variations (Guo et al. 2019; Träuble et al. 2021). Training without projectors delivers worse results compared to an initialization of a non-linear projector with a uniform distribution and freezing the parameters during pretraining (Appalaraju et al. 2020). This demonstrates that the learning capacity of the projector might not even be necessary, the non-linear transformation generated by the projector due to the additional layers and the separation of the pooled convolutional features from the final classification layer might already be beneficial (Appalaraju et al. 2020). Xue et al. (2024) identify reweighting as a key aspect of projectors and propose fixed reweighting projectors to generate improvements comparable to standard projectors. While most projectors consider the overall image and are task-agnostic, pixel-level CL applies specialized projectors, e.g. in DenseCL for dense prediction tasks (Hu et al. 2024b; Wang et al. 2021).

10.3.1 Guillotine regularization

Bordes et al. (2023a) name the practice of removing tailored projector layers Guillotine regularization (GR). The optimal layers to discard depend on the downstream task, the data and the training configuration (Bordes et al. 2023a). Similar ideas have already been discussed before, e.g. by Chen et al. (2020c), who fine-tuned from the middle layers of the projection head instead of abolishing the whole head, which increases performance and helps the model to profit more from deeper projectors. GR can mitigate misalignment of the pretext and downstream task due to insufficiently optimized networks, misaligned training objectives, or misalignment of train and test data distribution (Bordes et al. 2023a). It can improve generalization (Przewikeźlikowski et al. 2024), extract "augmentation invariant subspaces" and improve "transformation-related variability" (Yerxa et al. 2024). In SL, the

projector can actually be modified to control the trade-off between pretraining performance and transferability, as transferability increases with the number of hidden projector layers and their width (Sariyildiz et al. 2023). In CL, the optimal layer for feature extraction for the downstream task differs depending on how much the network suffers from pretext overfitting (Bordes et al. 2023a). The more the positive views are aligned with the downstream task, the closer the optimal cut should be to the last layer of the projector (Bordes et al. 2023a). While the feature quality for the downstream task can vary widely depending on the layer in the supervised setting, for SSL there seems to be a more consistent trends which layers are most suitable for different tasks (Bordes et al. 2023a). SimCLR pretrained on ImageNet has the best downstream performance when including either only the encoder (ImageNet, CIFAR10, Places205) or the first projector layer (Eurosat, ImageNet 1%, CLEVR) (Bordes et al. 2023a). The need for tailored GR demonstrates "the inability to design experimental setups and training criteria that learn structured and truly invariant representations with respect to an appropriate set of factors of variation" (Bordes et al. 2023a). Still, aligning pretext and downstream task and generalizable data augmentation should be more important than optimized GR (Bordes et al. 2023a).

10.3.2 Architecture

Non-linear projectors are more popular than linear ones as they can improve the performance by unlocking the expressive capabilities of the encoder (Hu et al. 2024b; Cosentino et al. 2022a). According to Chen et al. (2020b), a linear projector only improves their model performance by more than 3%, while a non-linear by more than 10%. Linear projectors can be restricting given complex relationships between views (Huang et al. 2021; Shwartz Ziv and LeCun 2024), but if they expand "the signal direction in the representations learned by encoders", they improve performance (Gui et al. 2023). Asymmetric projectors can lead to up to 5 percentage points improvement on TinyImageNet (Dubois et al. 2022). Ideally, one projector should be large while the other's architecture matches the downstream data (Dubois et al. 2022). More projector layers can bridge the gap between the performance difference using large and small batch sizes (Bordes et al. 2023b). Wider and deeper non-linear projectors result in less encoder limitation and better downstream performance (Cosentino et al. 2022a). Mialon et al. (2024) suggest choosing wider over deeper projectors, as high depths can have negative effects on the test accuracy, while resampling can decrease the sensibility regarding projector width. Wider projectors improve the encoder's representation learning and lower pairwise dependence (Balestriero et al. 2023; Guo et al. 2019; Mialon et al. 2024). They can handle stronger augmentations better (Ryali et al. 2021) and their converged parameters are closer to the initial ones (Mialon et al. 2024). Harder pretext tasks also require larger projectors (Ryali et al. 2021). CL with feature decorrelation profits from high-dimensional projector outputs (Balestriero et al. 2023; Bardes et al. 2022; Garrido et al. 2022; Zbontar et al. 2021). BT is more sensitive to projector dimensionality than BYOL or SimCLR regarding top-1 ImageNet accuracy, the best performance is achieved for the highest applied dimensionality (16384), while BYOL does not profit from deeper and wider projectors (Zbontar et al. 2021). For VICREg, 1024 dimensions are recommendable, as performance plateaus afterwards (Bardes et al. 2022). Consequently, large output dimensions are not always required (Balestriero et al. 2023; Garrido et al. 2022). The sensitivity towards the number of embedding dimensions can be decreased by an adequate projec-

tor architecture and good hyperparameter choices (Bardes et al. 2022; Garrido et al. 2022; Zbontar et al. 2021). The encoder size also has an influence: For linear projectors, combining a large encoder output layer and a small projector is favorable (Bordes et al. 2023c).

10.4 Encoder

The encoder typically is a ResNet or ViT model used for semantic feature extraction (Cosentino et al. 2022b; Ma et al. 2023; Ozbulak et al. 2023). Three types are common: image encoders, momentum encoders, and dictionaries (He et al. 2020; Rani et al. 2023). The encoder output is characterized by a lower entropy, better augmentation robustness and downstream task performance compared to the projector (Ma et al. 2023) but the use of a projector makes it harder to determine what has been learned in the encoder (Bordes et al. 2023b). The encoder representations is not directly affected by the loss function optimization (Xue et al. 2024). During training, "layer-wise progressive feature weighting" occurs where lower layers are characterized by "more normalized and less specialized representations" (Xue et al. 2024). Encoder representations in general profit significantly from SSL pretraining (Zhang et al. 2023c). The encoder size has a significant influence on the performance: Increasing the dimensionality of the last layer mitigates the pretraining bias and improves both SSL and SL (Bordes et al. 2023c). Top-1 accuracy for VICReg with ResNet-50 and ImageNet is improved by 7.6 percentage points when increasing the dimensions from 512 to 10240, and by additional 0.8 percentage points for 32768. For SimCLR, the increases are 8.0 and 1.4 percentage points, for BYOL 8.6 and 1.7 (Bordes et al. 2023c). CL usually includes a hidden uniform prior and a dimensional bottleneck, which is responsible for the implicit feature selection not being in accordance with the downstream task; thus increasing the dimensionality has a positive effect (Assran et al. 2022a; Bordes et al. 2023c). The improvements are generally larger for smaller and less powerful models (Bordes et al. 2023c). Increasing the encoder dimensions is beneficial for both balanced and unbalanced datasets as well as linear and non-linear projectors of any dimension (Bordes et al. 2023c). For ResNet, increasing the output dimension is more beneficial than increasing the width and depth of the overall network (Bordes et al. 2023c). According to Dubois et al. (2024), larger encoders improve the linear separability of the learned representation. The encoder design furthermore plays an important role for multimodal approaches, they can either be designed modality-specific (no fusion or late fusion) or multimodal (unified or late fusion) (Zong et al. 2024). CL approaches typically employ modality-specific encoders to learn embeddings whose similarity is computed via cross-modal dot products, which is easy but fails to capture complex multimodal interactions (Zong et al. 2024). Dual-encoders often use symmetric contrastive loss and allow scalable retrieval (Radford et al. 2021). Alternatively, "matching prediction can be carried out on a joint representation" which models "richer cross-modal interactions" (Zong et al. 2024). Both approaches can also be combined or the "embeddings from different encoders can be projected to a shared latent space" (Zong et al. 2024).

10.5 Relation of projector and encoder

As projector and encoder have different objectives, the performance of the last encoder layer and the projector layers are not always correlated (Bordes et al. 2023a). For both CIFAR10 and CIFAR100 with SimCLR trained for 200 epochs on ResNet-18 encoder, the

encoder representation has a significantly higher downstream correlation and lower error rate compared to the projection head representation (Ma et al. 2023). While the encoder rather learns distributed representations, the projector is more specialized regarding the pretext task (Yu et al. 2024). The projector optimizes the uniformity objective and thereby ensures that dissimilar samples have a maximum distance, while the encoder rather fosters alignment by minimizing the distance between positive views (Ma et al. 2023). The rank of the projected features is lower than the ranks of the last encoder layers, probably because the projector optimizes the contrastive loss (Gupta et al. 2022). This makes the projector features not easily generalizable for the downstream task, as information which has been encoded before is ignored (Gupta et al. 2022). The projector is responsible for a significant share of the invariance to augmentation methods and tends to be style-invariant, so encoder information is more versatile and general (Bordes et al. 2023a; Gupta et al. 2022). Unsuitable data augmentations degrade representation quality of the projector more compared to the encoder (Xue et al. 2024).

11 Generative learning

Generative approaches based on MIM consist of four major components: masking, encoder, RT, and head (Li et al. 2024a). The associated design choices are visualized in Fig. 10 and will be discussed in the following chapters.

11.1 Masking

Masking is important for MIM as it determines the learned abstraction level (Kong and Zhang 2023) and helps extracting local contextual information (Wang et al. 2022c). Consequently, the locality strength of MIM, which allows aggregating neighbor pixels in the ViT attention heads, is highly correlated with masking ratio and patch size (Xie et al. 2023a). Masking enables learning data-agnostic as well as occlusion-invariant features (Kong and Zhang 2023) and can decrease token redundancy (Choi et al. 2024). It can be understood as the general data augmentation strategy for MIM (Assran et al. 2022b; Kong and Zhang 2023). Due to the high token redundancy in CV, image token masking strategies differ from MLM (Kakogeorgiou et al. 2022).

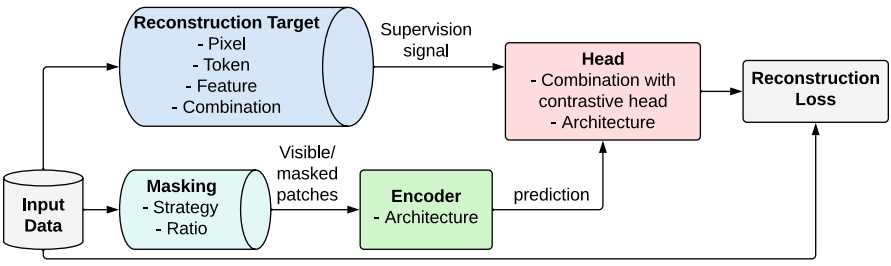

Fig. 10 Important design choices for MIM including the reconstruction target (pixel, token, feature, or their combination), masking (strategy and masking ratio) as well as the architecture of both encoder and head

11.1.1 Masking strategy

According to Li et al. (2024a), there are four main types of masking: Hard Sampling based on teacher-student frameworks, Mixture Masking where patches of diverse images are blended, Adversarial Masking based on adversarial learning, and Context Masking, where two patches from the same context are sampled. Hondru et al. (2025) distinguish between Random Masking and Informed Masking, which in turn can be divided into neural network based and feature based strategies. Although being one of the simplest approaches, Random Masking is overall the most popular strategy (Li et al. 2024a), especially for MIM based on raw and normalized pixels (He et al. 2022; Li et al. 2024a; Xie et al. 2022b; Zhang et al. 2024a). Random Masking can effectively create general-purpose pretext tasks (He et al. 2020). Given low-rank images, it preserves the original information while reducing computation resources (Cao et al. 2022). Still, Random Masking could be less useful for non-natural images, such as medical images (Xie et al. 2024). Blockwise Masking masks larger contiguous blocks instead of randomly scattered patches to remove structured information and force the model to understand a broader context (Bao et al. 2021). According to Dong et al. (2022), Blockwise Masking improves FT performance on feature-level but decreases it on pixel-level compared to Random Masking. Attention-guided Masking for distillation-based MIM with a student-teacher framework determines the masked parts using an attention mechanism and can therefore cover semantic entities within redundant images (Kakogeorgiou et al. 2022). Zhang et al. (2023d) developed attention-guided masking for object-centric MIM in remote sensing object detection to improve capturing small-scale objects. This offers a good balance between performance and computational cost (Li et al. 2024a). Multi-Masking can reduce the required batch sizes and training duration, as more information is extracted from the same amount of data, which demonstrates the interdependence of masking strategy, batch size and training duration (Baevski et al. 2023). Liu et al. (2023a) replace masked tokens from one image with visible tokens from another to facilitate MIM on hierarchical ViT. Joint Masking jointly selects nearby patches (Hu et al. 2022). Xie et al. (2023b) introduce frequency masking, a technique in which low-frequency image data is used to predict masked high-level visual information or vice versa. Shi et al. (2022) add a UNet based occlusion module that learns to mask semantically meaningful parts of the images adversarial to the feature encoder, which is superior to random masking, but requires high computational efforts (Chin et al. 2024). DPPMask aims to reduce the semantic change of the image due to masking (Xu et al. 2024). Low-level information can furthermore be included to improve masking (Hondru et al. 2025). Hinojosa et al. (2024) apply four filters to modify the uniform noise used for random masking, a strategy inspired by color noise used in image processing. Jarca et al. (2024) design masking as a curriculum learning problem and progressively increase the masking ratio, prioritizing patches with higher gradient magnitudes. Instead of designing and applying a masking strategy, some approaches attempt to learn an optimal masking strategy (Hondru et al. 2025): SemMAE (Li et al. 2022c) add semantic part learning, where attention maps capturing meaningful semantic image components are learned before masking. Kakogeorgiou et al. (2022) learn the masking strategy via a teacher network which uses the intact image to generate a mask for pretraining the student network. Wang et al. (2023f) apply a strategy based on direct feedback from the pretrained model, where those patches are masked that cause the highest reconstruction loss to focus on the most challenging regions during training.

11.1.2 Masking ratio

An optimal masking ratio prevents the model from predominantly focusing on low-level information, which can impede its capability to learn more complex, higher-level semantic features (Kong et al. 2023). Due to the higher redundancy in image data, MIM in general requires higher masking ratios than MLM (Li et al. 2024a; Zhang et al. 2023e). When training large segmentation models, large and diverse sets of masks improve generalization (Kirillov et al. 2023). MIM approaches based on low-level features often require higher masking ratios to learn abstract representations (Hondru et al. 2025). Table 15 summarizes important research regarding the optimal masking ratio taking the pretext task, the model architecture, the KT method, and the patch size into consideration. Masking ratio values that differ slightly from the column name are indicated via footnotes. CAE-v2 results are from Zhang et al. (2023a), SimMIM from Xie et al. (2022b), CIM from Fang et al. (2022), mc-BeiT from Li et al. (2022b), MVMAE from Ji et al. (2023), BIM from Luo et al. (2023) and MAE with ViT-L from He et al. (2022). EMAE and MAE results with ViT-B and LP are from Li et al. (2024b), while MAE results with ViT-B and FT are from Xu et al. (2024), who use an ImageNet subset with 100 classes. The data shows that masking ratios above 80% are not beneficial. Over-sized masks hinder high-level feature learning as "all observable descendants of a desirable high-level variable" are covered (Kong et al. 2023). While the model can still recover shape and texture of an object when key parts are moderately masked, masking over 90% leads to a mere prediction of the object's presence (Xie et al. 2022b). Over-sized masks reduce pretraining efficiency, increase pretraining duration and can cause inconsistencies (Li et al. 2024b). Very small masking ratios of 15% and below seem to be only beneficial for small networks or high patch ratios. Small mask ratios have minimal impact on the semantics of images, they cannot conceal critical information and acquire a sufficient level of pretext task difficulty (Li et al. 2024a). In combination with small patch sizes, small masking ratios result in low-level latent variables and prevent learning high-level features (Kong et al. 2023). Lower patch sizes in general result in more detailed representations which transfer worse (Xie et al. 2022b), which can be counterbalanced by increasing the masking ratio. Consequently, masking ratio and patch size should be optimized jointly. Most frameworks reach the best results around 50%, consequently medium masking ratios seem optimal in most cases. This is also corresponds to results from other domains, e.g. Xie et al. (2024) reach the best results for MedIM using 50% masking ratio for four different medical datasets. Still, the ideal masking ratio also depends on other design choices and on the input type (Chen et al. 2022a). Ratio and model size could be positively correlated, as networks with higher capacities should require more complex pretext tasks (Zhang et al. 2023a). High masking rates are more challenging (Xie et al. 2023d) and could prevent overfitting in big networks (Zhang et al. 2023a). Furthermore, increasing the masking ratio can speed up training (Li et al. 2023b). The masking strategy furthermore seems to matter: for Blockwise Masking with patch size 16, smaller networks tend to require smaller masking ratios. Both ViT-Tiny examples require 15%, ViT-S works best with either 15% or 25%, ViT-B requires predominantly 50%. Random Masking (60% on average for ViT-B) seems to require slightly higher masking ratios, presumably due to redundancies (Kakogeorgiou et al. 2022). Uniform Masking (25% for ViT-B) seems to work best with low ratios (Li et al. 2022d). Consequently, the ideal masking ratio seems to be highly dependent on the masking strategy.

11.2 Reconstruction targets

The RT influences the generation of supervision signals (Li et al. 2024a) and the downstream performance (Lu et al. 2023) of MIM models. Table 16 summarizes RT including examples, characteristics and contrastive elements. Hondru et al. (2025) distinguish between low-level and high-level target features, while the high-level target features can be further distinguished into models with a pretrained teacher or simultaneous training of teacher and student. MIM-based approaches use tokenizers, pixels, features (Li et al. 2024a) or a combinations of several RT (Jiang et al. 2022b). Xue et al. (2023) also enable MIM without reconstruction. Yang et al. (2023a) compare two different RT: raw pixels and high-level vision space which maps "patch features to a probabilistic distribution on a group of learnable weights". For LP, the high-level vision space improves accuracy by 9.4%, but for FT the difference is negligible (0.2%) (Yang et al. 2023a). Similarly, Ren et al. (2025) test both DeepMIM and MAE with different RT. MAE works best with DINO features or CLIP features (83.8% ImageNet top-1 accuracy), DeepMIM works best with CLIP features (84.8%) (Ren et al. 2025). Raw pixels generate the lowest performance (82.6% for MAE and 83.4% for DeepMIM) (Ren et al. 2025). Still, there seems to be no generally superior RT.

11.2.1 Pixel reconstruction

Pixels are the most fundamental reconstruction target (Li et al. 2024a). Targeting raw pixels (e.g. MAE and SimMIM) results in low computational overheads and a simple pretext pipeline, but those models tend to prioritize high-frequency details, which results in wasted modeling capacity (Liu et al. 2023c). They focus mostly on texture-rich details and might miss important foreground information (Liu et al. 2023d). Still, pixel-level pretext tasks have an outstanding performance in dense prediction downstream tasks (Xie et al. 2023a). iGPT treats each pixel as an individual token (Ren et al. 2024), also indicating the overlap of different reconstruction targets. SparK considers unmasked pixels as sparse voxels of point clouds (Tian et al. 2023). Pixel-based MIM approaches have been considered to be inferior to models with tokenizers, but recent improvements are closing this gap (Liu et al. 2023c), so they can be as effective as other targets (Gao et al. 2024; Ozbulak et al. 2023; Zhang et al. 2023e). Most pixel-based approaches include no contrastive elements.

11.2.2 Tokenizer

Tokens are mapping functions which represent images as dictionaries, they correspond to high-level semantic features of specific image segments (Li et al. 2024a) and help to exploit local patterns via token-level self-supervision (Park et al. 2023). Tokenizers vary in complexity and convert pixels into discrete tokens, masked tokens are then reconstructed by the decoder taking adjacent tokens into consideration, which is a non-differentiable process (Bao et al. 2021; Hondru et al. 2025; Park et al. 2023). Consequently, semantic-rich visual tokenizers promote MIM from pixel-level to semantic-level. If discrete tokens correspond well to the classes, these support the alignment of intra-class samples, which in turn increases the downstream performance (Du et al. 2024). In contrast, an inadequately designed tokenizer can obscure the distinctions between classes, resulting in performance

Table 15 Optimal mask ratio for ImageNet self-pretraining with patch size (P) (further information in Sect. 11.1.2)

Method	Model	P	KT	Masking ratio [%]											
				10	15	20	25	30	40	50	60	70	75	80	90
CAE-v2	ViT-T	16	FT		**77.8**		77.1			77.0			77.0		76.8
CAE-v2	ViT-T	16	LP		**69.3**		69.0			67.1			61.9		52.0
CAE-v2	ViT-S	16	FT		**82.7**		**82.7**			81.4			80.2		78.0
CAE-v2	ViT-S	16	LP		**77.5**		**77.5**			77.0			74.0		65.0
CIM	ViT-B	16	FT						82.8	**82.9**	82.8				
mc-BeiT	ViT-B	16	FT						**83.2**[a]		**83.2**		82.8		
CAE-v2	ViT-B	16	FT		85.2		**85.3**			85.3			84.3		82.1
CAE-v2	ViT-B	16	LP		80.1		80.5			**80.6**			79.4		71.6
MAE	ViT-B	16	FT						89.8	90.0	**90.1**	90.0	89.7		
EMAE	ViT-B	16	LP										**70.4**	68.1[b]	66.7[c]
CIM	ViT-B	16	FT						83.0	**83.3**	83.1				
mc-BeiT	ViT-B	16	FT						83.0[a]		83.1		**83.3**		83.2
MAE	ViT-B	16	FT						89.5	89.7	**89.9**	89.6		89.5	
MAE	ViT-B	16	LP										**64.4**	63.9[b]	60.7[c]
MVMAE	ViT-B	16	FT							79.2			79.1	**79.3**	79.2
RILS	ViT-B	16	FT								**82.7**		**82.7**		82.3
RILS	ViT-B	16	LP								**69.3**		68.5		67.6
SimMIM	Swin-B	16	FT						**82.7**		82.6			82.4	
SimMIM	Swin-B	32	FT						**82.7**		82.6			82.5	
SimMIM	Swin-B	4	FT						81.9		82.0			**82.1**	
SimMIM	Swin-B	8	FT						82.0		82.1			**82.4**	
SimMIM	Swin-B	16	FT						82.4		82.7			**82.8**	
SimMIM	Swin-B	32	FT						**82.9**		82.8			82.4	
SimMIM	Swin-B	32	FT	82.7		82.8		82.8	82.9	**83.0**	82.8	82.7		82.4	82.4
SimMIM	Swin-B	64	FT	**82.6**		**82.6**									
SimMIM	Swin-B	32	FT	**82.6**[d]			82.5		82.5[e]						
BIM	ViT-L	16	FT								81.8	83.8	**84.6**	84.3	82.7

Table 15 (continued)

Method	Model	P	KT	Masking ratio [%]											
				10	15	20	25	30	40	50	60	70	75	80	90
MAE	ViT-L	16	FT	83.4		83.4		84.7	84.9	**85.0**	84.9	84.9		84.5	83.0
MAE	ViT-L	16	LP	54.6		58.9		61.7	67.0	69.9	71.8	**73.5**		71.8	66.1

[a] 45%

[b] 85.7%

[c] 92.9%

[d] 11%

[e] 44%

Table 16 Different RT of MIM including the respective subtype, examples of specific models, CL elements and properties. The table is based on the work of Gui et al. (2024), Li et al. (2024a) and Ozbulak et al. (2023)

RT	Subtype	Examples	CL elements	Properties
High-level feature	Distillation	DMJD, SdAE	CLIP (DMJD)	Feature/ self-distillation
High-level feature	Offline teacher	MILAN, Img2vec	CLIP (MILAN)	Pretrained teacher
High-level feature	Online teacher	data2vec, data2vec-v2, dBOT	–	Semantic features
High-level feature	Fourier features	PixMIM, A2MIM, Ge2AE	Head (Ge2AE)	Fourier loss/masking
Low-level Feature	HOG features	MaskFeat, FastMIM	–	Local patterns
High-level feature	Patch features	ConMIM	Hybrid	Contrastive denoising
High-level feature	Position prediction	DILEMMA, DropPos	–	Postion encoding
Pixels	Pixel-based voxels	SparK	–	Sparse, hierarchical
Pixels	Pixel-based tokens	iGPT	–	Pixels as tokens
Pixels	Normalized pixels	MAE, SimMIM	–	Most basic target
Tokenizer	CLIP (offline)	BEiT-v2, BEiT-v3, CAE-v2	CLIP	Multimodal tokenizer
Tokenizer	dVAE (offline)	BEiT, CIM, mc-BEiT, PeCo	–	Low-level semantics
Tokenizer	iBot (online)	mc-BEiT	–	Self-distilled tokens
Tokenizer	VQGAN (offline)	LVM, MAGE, SeiT++	Head (MAGE)	GAN token learning
Combination	Multiple	MR-MAE, RC-MAE, MaskCLIP	Hybrid (MaskCLIP)	Hybrid
None	Student-teacher	MaskAlign	–	Visible patches
Hybrid	MAE target generator	DeepMIM	–	Applicable to various RT

decline (Du et al. 2024). In general, tokenization can have a significant impact on the performance. For BEiT, using visual tokens instead of predicting raw pixels is more influential than the masking strategy (Bao et al. 2021). iBot without a tokenizer leads to 29.8% LP accuracy on ImageNet-1K, while FT performance is 79.4%, consequently tokenization seems to be of considerable importance for LP (Zhou et al. 2022b).

Given that images are continuous and lack the "natural" token-structure of language, there are various approaches to tokenization (Zhang et al. 2023e; Zhou et al. 2022b). BEiT, mc-BEiT, and CIM use dVAE derived from DALL-E (Ramesh et al. 2021), which mainly captures low-level semantics and local details (Zhou et al. 2022b). With dVAE, the encoder transforms continuous embeddings into discrete tokens (Hondru et al. 2025). mc-BEiT applies dVAE for multiple-choice instead of single-choice by introducing soft-labels to avoid that patches with similar semantics have different token IDs (Li et al. 2022b). BeiT-v2, BeiT-v3, and CAE-v2 use CLIP (Radford et al. 2021), which is based on CL and can extract rich semantic information (Zhang et al. 2022a). CLIP includes multimodality knowledge, which can be beneficial for the downstream representations (Hou et al. 2022; Liu et al. 2023c, f). Both dVAE and CLIP require offline pretraining, which increases computational complexity (Liu et al. 2023d; Ozbulak et al. 2023). VQGAN is another offline tokenizer learning tokens via a vector-quantized GAN, which combines advantages of MIM and GAN (Lee et al. 2024). iBOT introduced their own online tokenizer inspired by BYOL

using a twin-teacher with cross-view similarity learning (Ozbulak et al. 2023; Zhou et al. 2022b).

11.2.3 Feature reconstruction

Features can either be low-level (Histograms of Oriented Gradients (HOG) Features, Position, Fourier and Patch Features) or high-level (offline and online teachers, distillation) (Li et al. 2024a). While Img2vec and MILAN employ an offline teacher requiring additional pretraining, online teacher models such as data2vec, data2vec-v2, and dBOT either combine multiple modalities or a multi-stage distillation scheme (Liu et al. 2022c). Semantic-rich teachers use features as reconstruction targets and thereby improve both performance and capture rich semantic correlations across regions (Xue et al. 2023). Instead of reconstructing pixels or tokens, student models are trained to align their internal feature representations with those of the teacher (Wang and Yoon 2022). Methods that align student and teacher models efficiently without requiring explicit pixel or token reconstruction could reduce the computational cost, but come with a complex pipeline (Wang and Yoon 2022). Feature distillation reaches comparable performance across diverse teachers, especially after multi-stage distillation, indicating that the improvements stem more from the asymmetric architecture than the distillation itself (Shi et al. 2023b). Furthermore, feature reconstruction methods use multiple elements of CL, demonstrating how different learning paradigms cross-fertilize and intermingle.

11.3 Encoder

The main purpose of the encoder in MIM is the prediction of the masked vector by processing the visible patches (Hondru et al. 2025; Li et al. 2024a). Some encoders additionally use the masked patches (Li et al. 2024a). Although the encoder can have different architectures, the introduction of ViT encoders was essential for the development of MIM, as their stable representation propagation facilitates the generation of reliable reconstruction results (Cao et al. 2022). As CNN operate on small receptive fields, they need to be considerably deeper than ViT and rather reinforce the locality bias of MIM (Cao et al. 2022; Xie et al. 2023a). ViT have a more global approach, which e.g. might allow MAE to reconstruct images based on few unmasked patches (Cao et al. 2022). Still, some research focuses on the on compatibility of MIM with CNN encoders (see e.g. Fang et al. (2022); Li et al. (2023d); Woo et al. (2023)). The different ViT encoder-blocks have different purposes (Alkin et al. 2025). Early blocks rather learn general features, which is beneficial for reconstruction loss and representation quality (Alkin et al. 2025). Intermediate ones foster abstraction, which is more favorable for feature quality, while later blocks focus more on reconstruction loss improvement, which is detrimental for the feature quality (Alkin et al. 2025). Additionally, bigger encoders increase the representation's linear separability (Dubois et al. 2024). The encoder design is crucial for multimodal MIM. Cross-modal alignment can be fusion free resulting in a simpler architecture and easier scaling, but cannot model all multimodal interactions, resulting in a poorer performance on downstream tasks such as visual reasoning (Zong et al. 2024). Alternatives are late fusion (e.g. applying cross-attention), which requires more computational resources, or unified encoders, which can process tokens from different modalities

and are less sensible regarding missing modalities, but can be inefficient regarding retrieval tasks (Zong et al. 2024).

11.4 Head

MIM heads use the latent feature representation of the encoder to reproduce the masked areas (Xie et al. 2022b) and predict the reconstruction loss (Li et al. 2024a). The choice of an appropriate head depends on the reconstruction target, most frameworks either use an MIM decoder, a contrastive head, or both (Li et al. 2024a). Similar to projectors in CL, heads are discarded after pretraining to prevent the backbone features from overfitting to the pretraining objective, thereby improving the transfer capability (El-Nouby et al. 2024; Chen et al. 2021c, 2020b; Caron et al. 2021a, b; Grill et al. 2020) and enabling the encoder to learn high-level features required for the downstream task (El-Nouby et al. 2024). MIM heads can be linear (e.g. BEiT, BeiT-v2, data2vec), MLP, or ViT, which is the most popular option (Li et al. 2024a). They can also be combined with contrastive heads, e.g. by including additional contrastive loss (Li et al. 2024a; Mishra et al. 2022; Yi et al. 2023; Zhou et al. 2023c). The inclusion of contrastive heads can boost performance (Li et al. 2024a). Light-weighted heads, even those consisting simply of a linear layer, allow a performance comparable to more complex heads while reducing training time (Almalki and Latecki 2023; Fang et al. 2022; Liu et al. 2024a; Xie et al. 2022b). When using a simple linear head, masked tokens should be included in the encoder input, so these can interact with the visible patches and enable effective reconstruction (Li et al. 2024a). Without the presence of these tokens, more complex architectures, such as transformers, are required (Li et al. 2024a). Xie et al. (2022b) compare linear, MLP and Swin heads and find that the linear heads performs best regarding ImageNet-1K FT performance; demonstrating that the strongest reconstruction capabilities in MIM do not necessarily result in the highest downstream performance. While more complex heads and higher target resolutions enhance pretraining performance, the benefits do not automatically transfer to the downstream task (Xie et al. 2022b). Replacing an MLP head with a full-fledged ViT, while keeping the depth and width the same, can only provide minor improvements but increases computational cost significantly (El-Nouby et al. 2024). As heads can be discarded after pretraining, they can determine which layer the downstream features are extracted from. The highest representation quality can be achieved when the features are extracted from the middle of the network, as detailed pixel prediction or reconstruction is not beneficial for downstream classifications (Chen et al. 2020d; El-Nouby et al. 2024). The generative pretext objective does not necessary lead to the last layers having the highest semantic content (El-Nouby et al. 2024).

12 Knowledge transfer

As the KT between the pretext and downstream task has a considerable influence on the quality, evaluation, and usability of the transferred representations as well as their effective reuse for downstream training (Goyal et al. 2019; Kolesnikov et al. 2019; Misra and Maaten 2020; Zhang et al. 2023c), KT is an important design choice. FT and LP are the most common types of KT. They are illustrated in Fig. 11. There is no inherent correlation between LP and FT performance, so the performance in one transfer setting might not be generaliz-

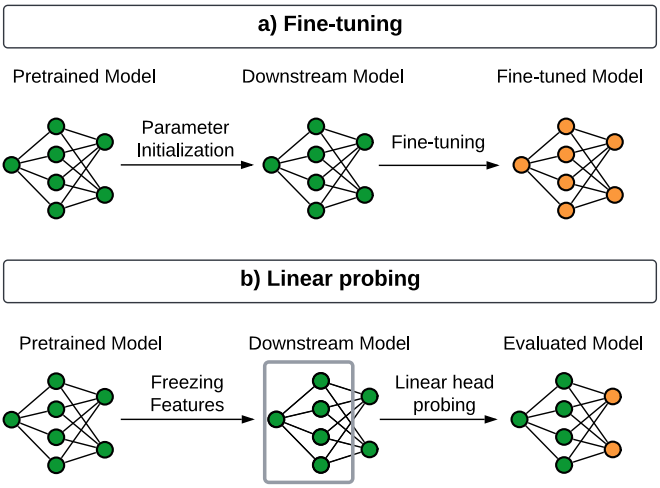

Fig. 11 Schematic overview of FT and LP as the two main types of KT. FT employs the pretraining for parameter initialization and retrains the network on the downstream task and objective. In contrast, LP freezes the features after pretraining and only trains the last layer on the downstream task and objective

able (Chen and He 2021; He et al. 2022; Marks et al. 2025; Newell and Deng 2020; Ozbulak et al. 2023). Instead, the two approaches can be employed to assess two distinct criteria. LP rather assesses the quality of the representation, whereas FT evaluates the pretext transferability (He et al. 2023). Still, the specific KT settings and their performance varies widely across SSL methods (Lee et al. 2023b). The choice between FT and LP is of considerable importance (Ericsson et al. 2022b, 2021; Kolesnikov et al. 2019) and can have a higher impact on performance than the model size (Chen et al. 2021c).

12.1 Linear probing

During LP, all trainable parameters besides the last linear layer are frozen, which is trained on the downstream data, while the lower layers remain unchanged (Kumar et al. 2022a). The goal is to learn predictions from features which have been extracted in the pretext task, so the pretrained model is used as a feature extractor (Ericsson et al. 2022b; Misra and Maaten 2020; Ozbulak et al. 2023). LP requires sufficient dimensionality and equivalence classes that can be classified linearly for optimal representations (Dubois et al. 2022). Non-linear transfer techniques are superior given difficult domain adaptations, while LP is superior in easier scenarios, which could profit more from features extracted from deeper layers (Wu et al. 2024). Compared to FT, LP generates more comparable results due to the fewer parameters having to be tuned (Ericsson et al. 2022a), consequently it is used in numerous publications aiming to prove the performance of their SSL approach. Still, ranking models based on LP cannot be generalized to the FT performance (Newell and Deng 2020). While LP is used for evaluation, performance optimization (Balestriero et al. 2023) and hyperparameter testing (Bao et al. 2021), it cannot make use of non-linear features (Chen et al. 2024a; He et al. 2022; Yi et al. 2023). In turn, the degree of linear transferability is high when data from one class and different domains is more related to itself than to data from other classes (Hao-Chen et al. 2022b; Marks et al. 2025). LP tests whether the pretext task generates repre-

sentations with linearly separable groups in the latent space (Scheibenreif et al. 2022) and works best on high-level pretext representations (Assran et al. 2023). The downstream task can benefit from reduced sample complexity thanks to learning good representations with LP (Arora et al. 2019). Besides, LP is in general beneficial if the downstream task is trained in a sparse data regime, as overfitting can rather be avoided due to the limited number of trainable parameters (Cole et al. 2022; Ericsson et al. 2022b; Dosovitskiy et al. 2015; Gao et al. 2020). Data and parameter efficiency are better with LP, while FT is characterized by a higher transfer performance (Jiang et al. 2022a). Both have a comparable modality and task scalability (Jiang et al. 2022a). Pretext models trained on coarse-grained data are not good at discriminating fine-grained downstream classes when using LP (Heim 2023). Lee et al. (2023b) compare hyperparameter-sensitivity of LP of three contrastive (MoCov3, DINO, Mugs) and three MIM-based methods (MAE, iBot, MSN). Hyperparameters e.g. include epochs, batch sizes, optimizer, learning rate, weight decay, patch tokens, and BN layers (Lee et al. 2023b). MAE varies most with nearly 35 percentage points difference between the best and the worst setting for ImageNet on ViT-B/16 (Lee et al. 2023b). In general, LP seems to be very sensitive to hyperparamter settings, which can be reduced using input normalization such as batch normalization or embedding normalization (Marks et al. 2025; Lee et al. 2023b).

12.2 Fine-tuning

In contrast to LP, FT retrains the entire model on the downstream data and objective after pretraining (Misra and Maaten 2020; Ozbulak et al. 2023). Consequently, FT is a form of domain adaptation minimizing domain mismatches (Pan et al. 2023a). The initialization has an implicit regularization effect (Kumar et al. 2022a; Neyshabur et al. 2021). Ideally, FT should both preserve the helpful information from pretraining and adapt to the downstream objective (Lee et al. 2023a). Unlike LP, FT can tune both the feature extractor and the prediction head to the downstream objective and could suffer from a pretraining bias (Kumar et al. 2022a). While being computationally expensive, FT is characterized by a stronger transfer learning performance (Balestriero et al. 2023) and is the dominant KT approach for SSL models (Han et al. 2021). According to Zhang et al. (2023c), there are three common FT types: Fixed FT (Goyal et al. 2019), where the encoder is frozen and only the decoder stack or the fully-connected layers are fine-tuned, warm-up FT (Taleb et al. 2000), where the whole network is frozen for few epochs, then full FT is applied (Zhang et al. 2023c) and full FT (or end-to-end FT), where all weights are fine-tuned. Full FT is the most common approach in many domains, such as medical imaging (Khan and Fang 2023; Zhang et al. 2023c) and is most beneficial in regimes with enough data to update all parameters, so large labeled downstream datasets are to be preferred (Chen et al. 2020c; Chi et al. 2021; Devlin et al. 2018; Ericsson et al. 2022b; Hu et al. 2020; Jiang et al. 2022a; Pathak et al. 2016). Still, there is little insight into whether this strategy is the best way to leverage pretext knowledge, especially given that different SSL paradigms focus on different features (Khan and Fang 2023).

Zhang et al. (2021d) compare different forms of FT: while the best approach (Core-tuning) has 92.73% accuracy on ImageNet20, the worst (standard CE-based FT) reaches 4.45 percentage points less. Consequently, different FT strategies have a huge impact on the performance (Zhang et al. 2021d). Furthermore, FT hyperparameter choices (such as batch

size, epochs or learning rate) can change significantly across frameworks and are often over-adapted to the downstream dataset (Lee et al. 2023b). Different frameworks show different sensitivities regarding the FT protocol, which seems to be influenced by the network size (Zhou et al. 2022b). Many SSL approaches require expertise and task-specific design choices and an excellent understanding of the training behavior (Azizi et al. 2023). Most common FT approaches have been designed for SL and prevent the downstream task to differ too much from the pretrained version, which can result in negative transfer (Chen et al. 2019a) if the downstream and the pretext task differ significantly (Zhang et al. 2021d). Adaptions to SSL include improved model initialization (Pan et al. 2023a), such as injecting noise into the initialized parameters before FT to decrease the pretraining bias and thereby the dataset gap (Wu et al. 2022). Random initialization of the fully-connected layers (RIFLE) (Li et al. 2020) can improve learning low-level features. Prototypical FT with non-parametric nearest-neighbor classification (Medina et al. 2020) and prototypical initialization (Triantafillou et al. 2020) can enhance results in few-shot scenarios (Medina et al. 2020). Surgical FT or layer-wise FT only fine-tune specific layers to enhance the adaption to distribution shifts (Lee et al. 2023a). Shallow FT fine-tunes a shallower network consisting of the first three quarters of the original pretrained networks (layers 1-9 for MoCo and MAE) (Khan and Fang 2023). The last quarter of the network seems to be particularly difficult to adapt to the respective downstream task, fine-tuning them can significantly reduce the overall performance (Khan and Fang 2023). This can be attributed to the fact that the final layers could be too specialized for the pretext task, which prevents them from capturing more general representations that benefit downstream tasks (Khan and Fang 2023). Shallow FT outperforms both full and surgical FT for MoCov3 and MAE with ViT-B and can improve accuracy on ImageNet by up to 5.48% (Khan and Fang 2023). Given ViT and small datasets, fine-tuning only the weights corresponding to the multi-headed self-attention residual block layers and freezing the feedforward network can exceed the results of full FT, as restricting the number of trainable weights probably has a regularizing effect (Touvron et al. 2022). In contrast, full FT is rather beneficial for larger datasets and larger versions of ViT (Touvron et al. 2022).

The appropriate FT strategy can differ depending on the learning paradigm. For MIM, concept-wise FT should improve feature representations on patch level by maximizing MI between samples from the same category and front-door adjustment (Yang et al. 2023b). Khan and Fang (2023) show that fine-tuning the second quarter of the network can be superior for CL, while fine-tuning the third quarter is best for MIM-based SSL. This is the case across a wide range of downstream data for for medical images with sizes from 100 to 75.000 samples (Khan and Fang 2023). In CL, discrepancies can arise due to an altered image scale caused by data augmentation, which can be mitigated by pretraining at a lower resolution and subsequently increasing it for FT (Touvron et al. 2022). SCL adds contrastive loss to FT (Gunel et al. 2020), which has both a regularizing and optimizing effect (Zhang et al. 2021d). Contrastive Initialization (COIN) adds a class-aware initialization stage to improve standard FT for CL (Pan et al. 2023a). COIN is supposed to compensate for the fact that two samples from the same class can be considered as far apart in the learned feature space during pretraining (Pan et al. 2023a). Compared to SCL, COIN increases accuracy by 5.0% on ImageNet-20 and 2.6% on CIFAR-100 (Pan et al. 2023a). Contrast-regularized FT mines hard negative pairs and applies contrastive loss during FT, aiming to smooth the decision boundary of the classifier (Zhang et al. 2021d). Goyal et al. (2023) propose contrastive

FT, where downstream labels are employed as prompts, thereby enabling the minimization of the contrastive loss between image and prompt embeddings during FT. These enhancements demonstrate the importance of FT and the drive for further optimization. However, this paper will mostly focus on standard full FT.

12.3 Performance comparison

Given the specific characteristics of LP and FT, which form of KT is superior in which application scenario? Gui et al. (2024) compare 26 SSL frameworks with predictive, contrastive or generative pretraining on ImageNet and test both methods on segmentation (Pascal VOC 2007 COCO and ADE20K) as well as detection (Pascal VOC 2007, Pascal VOC 2007+12 and COCO). FT results are superior to LP results for every tested SSL framework (Gui et al. 2024). Regarding LP, CL approaches perform best, probably due "to the algorithm's ability to generate well-structured latent spaces, wherein distinct categories are effectively separated, and similar categories are appropriate" (Gui et al. 2024). CL could be more suitable for LP as the discriminative loss function fosters linearly separable features (Wei et al. 2022a). MIM can in turn achieve superior performance given FT, probably thanks to the expressive non-linear features (He et al. 2022), while CL has a tendency for overfitting (Gui et al. 2024; Wang and Qi 2023; Robinson et al. 2021b; Wei et al. 2022b).

Tables 17 and 18 compare FT and LP performance on ImageNet and supports these findings. All models have a considerably higher FT than LP performance (on average 14.9% gain from FT compared to LP). The average gain of FT is only 7.8% for contrastive frameworks (see Table 17), but 20.0% for generative and 13.1% for hybrid (see Table 18). The

Table 17 Performance of contrastive ImageNet-1k self-pretraining with LP and FT depending on pretext task, model architecture and pretraining epochs including the improvement due to FT compared to LP (Gain). All results are from the original publications besides the DINO results with 3200 and 1600 epochs (Zhou et al. 2022c), MoCo-v3 with 600 epochs (Chen et al. 2024a), SimCLR (Mishra et al. 2022), SwAV (Caron et al. 2021b) and MoCo-v2 with ViT-H (Ozbulak et al. 2023)

Method	Architecture	Epochs	LP	FT	Gain
BYOL	ResNet-50	1000	74.3	77.7	4.6
DINO	ViT-S	300	77.0	81.5	5.8
DINO	ViT-B	300	78.2	82.8	5.9
DINO	ViT-S	3200	77.0	82.0	6.5
DINO	ViT-B	1600	78.2	83.6	6.9
MoCo	ResNet-50	200	60.6	77.3	27.6
MoCo-v2	ResNet-50	800	71.1	75.5	6.2
MoCo-v3	ViT-B	300	76.7	83.2	8.5
MoCo-v3	ViT-L	300	77.6	84.1	8.4
MoCo-v3	ViT-S	600	73.1	81.7	11.8
MoCo-v3	ViT-B	600	76.2	83.0	8.9
Mugs	ViT-S	3200	78.9	82.6	4.7
Mugs	ViT-B	1600	80.6	84.3	4.6
SimCLR	ViT-L	800	73.9	83.4	12.9
SimCLR-v2	ResNet-50	400	71.7	76.3	6.4
SimCLR-v2	ResNet-50 (2x)	400	75.6	79.1	4.6
SimCLR-v2	ResNet-101	400	73.6	78.2	6.3
SimCLR-v2	ResNet-101 (2x)	400	77.0	80.7	4.8
SwaV	ResNet-50	800	75.3	77.8	3.3

Table 18 Performance of generative and hybrid ImageNet-1k self-pretraining with LP and FT depending on pretext task, model architecture and pretraining epochs, including the improvement due to FT compared to LP (Gain). All results besides iBOT with ViT-S or ViT-B (Zhou et al. 2022c) as well as iBot with ViT-H and MaskFeat (Ozbulak et al. 2023) are from the original publications

G/H	Method	Architecture	Epochs	LP	FT	Gain
G	BEiT	Vit-B	800	56.7	83.2	46.7
G	BEiT	Vit-L	300	73.5	85.2	15.9
G	BEiT-v2	ViT-B	300	80.1	85.0	6.1
G	BigBiGAN	ResNet-50	800	55.4	76.3	37.7
G	BigBiGAN	RevNet-50 (4x)	800	60.8	76.6	26.0
G	BootMAE	ViT-B	300	64.1	84.0	23.7
G	CAE	ViT-S	300	51.8	82.0	58.3
G	CAE	ViT-B	1600	70.4	83.9	19.2
G	CAE	Vit-L	1600	78.1	86.3	10.5
G	CAE	Vit-B	300	64.1	83.6	30.4
G	CAE-v2	ViT-S	300	77.5	83.1	7.2
G	CAE-v2	ViT-B	300	80.7	85.5	5.9
G	CAE-v2	ViT-T	300	69.3	77.8	12.3
G	iBOT	ViT-S	3200	77.9	82.3	5.6
G	iBOT	ViT-B	1600	79.5	83.8	5.4
G	iBOT	ViT-H	1000	81.0	88.3	9.0
G	iGPT	GPT-L	100	65.2	72.6	11.3
G	M-MAE	ViT-B	200	62.4	83.1	33.2
G	M-MAE	Vit-L	200	66.0	84.3	27.7
G	MAE	Vit-B	1600	68.0	83.6	22.9
G	MAE	Vit-L	1600	75.8	85.9	13.3
G	MAE	Vit-L	800	75.1	85.1	13.3
G	MAE	ViT-H	1600	76.6	86.9	13.4
G	MaPeT	ViT-B	300	73.5	83.6	12.1
G	MaskFeat	ViT-L	1600	67.7	84.0	24.1
G	MILAN	ViT-B	400	79.9	85.4	6.9
G	MILAN	ViT-L	400	84.3	87.8	4.2
G	MSG-MAE	ViT-B	300	75.6	85.3	11.4
G	SdAE	ViT-B	300	64.9	84.1	29.6
G	SiameseIM	ViT-B	1600	78.0	84.1	7.8
G	SparK	ConvX-B	1600	54.7	84.8	55.0
H	CAN	ViT-B	800	74.0	83.4	12.7
H	CAN	Vit-L	800	76.2	84.7	11.2
H	CMAE	ConViT-B	1600	73.9	85.3	15.4

lowest difference of LP and FT is 2.3% (SwaV), the highest 58.3% (CAE on ViT-S). iBOT has the best overall FT accuracy (88.3%) of all analyzed approaches, MILAN has the highest LP accuracy (84.3%) and the second highest FT result (87.8%). All scenarios with a gain higher than 15% are generative. The models with the lowest LP performance (below 70%) are all generative and have the highest gains, demonstrating that the high gains of MIM models rather stem from bad LP performances than extraordinary FT performances. One explanation is the tendency of MIM to learn a bias towards texture, while CL methods are rather biased regarding shape, which could be more beneficial for LP (Park et al. 2023). Additionally, data augmentation may enhance the capacity of CL methods to comprehend

representation consistency, thereby supporting the applicability of LP (Tian et al. 2022). The experiments conducted by Tian et al. (2022) indicate that the incompatibility of MIM and LP is primarily due to the difficulty of MIM in reconstructing semantic features. The lower semantic quality impedes the efficacy of LP or other "off-the-shelf" approaches (Assran et al. 2023). Token-based MIM usually requires FT to sufficiently learn semantics (Tian et al. 2022) while the linear head of LP is ill-suited to exploit pretext features distributed across different tokens (Tian et al. 2022). Nevertheless, further investigation is required to determine whether the generative pretext task itself leads to a lack of suitability for LP or whether ViT backbones are disadvantageous for LP, which are more widely used for MIM (Marks et al. 2025). Interestingly, MIM with contrastive elements or hybrid models (e.g. CAE) are not characterized by a better LP performance compared to MIM without contrastive elements (e.g. MAE). Still, CLIP seems to increase LP performance considerably compared to dVAE (Ozbulak et al. 2023). Longer pretraining on larger and more diverse datasets can enhance the performance of LP on MIM (Assran et al. 2023), but it is no universal solution. While LP performance of CAE profits significantly from increasing pretraining duration (70.4% instead of 64.1% when training with 1600 instead of 300 epochs with CAE and ViT-B), the difference for MAE is not significant. Furthermore, the differences can be huge within one framework according to the network size: While CAE has the highest gain of all models on ViT-S (58.3%), the difference is considerably lower on ViT-L (10.5%). Increasing the transformer size nearly always decreases the gain for the same pretext task (besides when MAE Vit-L is upgraded to ViT-H and iBOT from ViT-B to ViT-H), which in turn means that LP profits more from bigger network sizes than FT does. Consequently, scaling up the size of the network can be seen as an important strategy for improving LP. In general, the performance difference of two similar pretext tasks can be huge. While MoCo has a gain of 27.6%, MoCo-v2 only has 6.2% (using ResNet-50), which could be attributed to the different data augmentation strategies. BeiT on Vit-B has 46.7%, while BeiT-v2 has only 6.1%, which could be attributed to the different tokens used. However, these self-pretraining results with ImageNet cannot necessarily be generalized to other datasets, as KT is highly data-sensitive. For SimCLR, the gain due to FT varies with dataset: using iNat21, LP only reaches about 40%, while FT reaches about 70% top-1 accuracy (Cole et al. 2022). In contrast, the gain is below 10 percentage points for ImageNet and below 5 for Places 365 (Cole et al. 2022). For CIFAR-10 with iGPT-L, LP or FT only makes a moderate difference (96.3% vs. 99.0%) (Chen et al. 2020d). Therefore, Sect. 12.4 takes a closer look at the relationship between KT and domain gap.

12.4 Domain gaps and knowledge transfer

In general, SSL is less sensitive to domain gaps compared to supervised transfer learning (Yang et al. 2020a). However, due to the different characteristics of FT and LP, their suitability for training SSL models with domain gaps might differ. Table 19 summarizes the accuracy gain due to FT compared to LP given different domain gaps for both contrastive and generative SSL models that have been pretrained on ImageNet data. All downstream datasets are significantly smaller than ImageNet. The domain gap is measured via Maximum Mean Discrepancy (MMD), calculated by Chen et al. (2022b). On average, FT is 14.0% better. For MIM, FT always generates better results than LP, which could be caused by the low linear separability of representations learned with generative approaches while CL ensures

Table 19 Relative improvement (gain) of FT compared to LP depending on the domain gap of ImageNet pretraining with contrastive SSL (C) or generative MIM (G) including the pretext tasks, model architecture and the different downstream datasets. The domain gap is measured via MMD by Chen et al. (2022b). The results show that FT is always superior for MMD = 0 and nearly always better for MMD > 1. ReLIC-v2 results are from Tomasev et al. (2022) and TWIST results from Giakoumoglou et al. (2025). All BT, EVA, MILAN, MoCo-v3 and PixMix results are from Marks et al. (2025), BEiT-v2 from Peng et al. (2022), DINO from Caron et al. (2021b), M-MAE from Tan et al. (2024), iBot from Dimitrovski et al. (2024) and MAE from Gao et al. (2022). BYOL with MMD = 0.00 and MMD = 1.62 are from Marks et al. (2025). All MMD = 0.63 results are from Ericsson et al. (2021) besides the SimCLR result from Chen et al. (2020b). The MMD = 0.74 results are from Sirotkin et al. (2024) besides the first BYOL, the second SimCLR (Grill et al. 2020) as and the second SimCLR result (Chen et al. 2020b). All MMD = 1.64 results are from Ericsson et al. (2021) besides the second BYOL and the SimCLR result (Grill et al. 2020). MMD = 2.27 results are from Sirotkin et al. (2024) besides the first BYOL and the SimCLR result (Grill et al. 2020). For MMD = 2.39, the first BYOL, MoCo, MoCo-v2 and SwaV results are from Ericsson et al. (2021), while the second BYOL result from Marks et al. (2025) and the SimCLR result from Chen et al. (2020b). All MMD = 2.95 results are from Sirotkin et al. (2024) besides SimCLR from Grill et al. (2020)

C/G	Method	Model	Down data	MMD	LP	FT	Gain
C	BT	ResNet-50	ImageNet-1K	0.00	71.5	74.2	3.6
G	BEiT-v2	ViT-B	ImageNet-1K	0.00	80.1	85.0	6.1
C	BYOL	ResNet-50	ImageNet-1K	0.00	64.0	75.6	15.3
C	DINO	ViT-B	ImageNet-1K	0.00	78.2	82.8	6.9
G	EVA	ViT-B	ImageNet-1K	0.00	60.5	82.1	26.3
G	MILAN	ViT-B	ImageNet-1K	0.00	76.4	84.4	9.5
G	M-MAE	ViT-B	ImageNet-1K	0.00	62.4	83.1	33.2
C	MoCo-v3	ResNet-50	ImageNet-1K	0.00	73.9	80.8	8.5
G	PixMix	ViT-B	ImageNet-1K	0.00	49.5	79.2	37.5
C	BYOL	ResNet-50	Caltech-101	0.63	91.5	89.4	**−2.3**
C	DeepCluster-v2	ResNet-50	Caltech-101	0.63	91.3	90.8	**−0.6**
C	MoCo	ResNet-50	Caltech-101	0.63	75.3	75.0	**−0.4**
C	MoCo-v2	ResNet-50	Caltech-101	0.63	87.9	84.4	**−4.1**
C	ReLIC-v2	ResNet-50	Caltech-101	0.63	92.8	93.2	0.4
C	SimCLR	ResNet-50	Caltech-101	0.63	90.3	92.1	2.0
C	SwAV	ResNet-50	Caltech-101	0.63	90.8	89.9	**−1.0**
C	TWIST	ResNet-50	Caltech-101	0.63	94.5	93.5	**−1.1**
C	BYOL	ResNet-50	DTD	0.74	75.5	76.2	0.9
C	BYOL	ResNet-50	DTD	0.74	76.9	73.6	**−4.5**
C	DeepCluster-v2	ResNet-50	DTD	0.74	78.6	74.8	**−5.1**
C	MoCo	ResNet-50	DTD	0.74	68.8	65.4	**−5.2**
C	MoCo-v2	ResNet-50	DTD	0.74	73.9	69.5	**−6.3**
C	SimCLR	ResNet-50	DTD	0.74	74.5	73.2	**−1.8**
C	SimCLR	ResNet-50	DTD	0.74	75.7	75.4	**−0.4**
C	BT	ResNet-50	CUB	1.62	71.4	86.8	17.7
C	BYOL	ResNet-50	CUB	1.62	71.0	93.5	24.1
G	EVA	ViT-B	CUB	1.62	27.2	89.8	69.7
G	MILAN	ViT-B	CUB	1.62	84.6	96.5	12.3
C	MoCo-v3	ResNet-50	CUB	1.62	72.8	86.9	16.2
G	PixMix	ViT-B	CUB	1.62	23.4	80.0	70.8
C	BYOL	ResNet-50	Flowers	1.64	94.5	94.5	0.0
C	BYOL	ResNet-50	Flowers	1.64	96.1	97.0	0.9
C	DeepCluster-v2	ResNet-50	Flowers	1.64	94.7	95.3	0.6
C	MoCo	ResNet-50	Flowers	1.64	82.1	89.5	8.3
C	MoCo-v2	ResNet-50	Flowers	1.64	90.1	94.4	4.6

Table 19 (continued)

C/G	Method	Model	Down data	MMD	LP	FT	Gain
C	ReLIC-v2	ResNet-50	Flowers	1.64	95.6	97.9	2.3
C	SimCLR	ResNet-50	Flowers	1.64	92.6	96.6	4.1
C	SwAV	ResNet-50	Flowers	1.64	94.6	95.5	0.9
C	TWIST	ResNet-50	Flowers	1.64	93.4	97.1	3.8
G	iBot	ViT-B	UCM	1.89	96.0	97.4	1.4
G	MAE	ViT-B	UCM (80%)	1.89	92.9	99.2	6.4
G	MAE	ViT-B	UCM (50%)	1.89	84.5	97.9	13.7
C	BYOL	ResNet-50	CIFAR-10	2.27	91.3	97.8	6.6
C	BYOL	ResNet-50	CIFAR-10	2.27	93.3	97.0	3.8
C	DeepCluster-v2	ResNet-50	CIFAR-10	2.27	94.0	97.1	3.2
C	MoCo	ResNet-50	CIFAR-10	2.27	80.2	93.9	14.6
C	MoCo-v2	ResNet-50	CIFAR-10	2.27	92.3	96.5	4.4
C	ReLIC-v2	ResNet-50	CIFAR-10	2.27	95.6	97.9	5.3
C	SimCLR	ResNet-50	CIFAR-10	2.27	90.5	97.4	7.1
C	TWIST	ResNet-50	CIFAR-10	2.27	74.4	97.9	31.6
C	BT	ResNet-50	CIFAR-100	2.39	57.1	84.1	32.1
C	BYOL	ResNet-50	CIFAR-100	2.39	77.9	84.0	7.3
C	BYOL	ResNet-50	CIFAR-100	2.39	54.2	71.0	23.7
G	EVA	ViT-B	CIFAR-100	2.39	54.6	87.4	37.5
G	MILAN	ViT-B	CIFAR-100	2.39	70.9	90.0	21.2
C	MoCo	ResNet-50	CIFAR-100	2.39	57.7	71.5	19.3
C	MoCo-v2	ResNet-50	CIFAR-100	2.39	74.9	71.3	**−5.0**
C	MoCo-v3	ResNet-50	CIFAR-100	2.39	76.5	79.6	3.9
G	PixMix	ViT-B	CIFAR-100	2.39	50.0	81.1	38.3
C	ReLIC-v2	ResNet-50	CIFAR-100	2.39	78.2	85.3	8.3
C	SimCLR	ResNet-50	CIFAR-100	2.39	71.6	85.9	16.6
C	SwAV	ResNet-50	CIFAR-100	2.39	79.6	84.4	5.7
C	TWIST	ResNet-50	CIFAR-100	2.39	66.8	86.5	29.5
C	BYOL	ResNet-50	Aircraft	2.95	53.9	79.5	32.2
C	DeepCluster-v2	ResNet-50	Aircraft	2.95	54.5	82.5	33.9
C	MoCo	ResNet-50	Aircraft	2.95	35.6	75.6	52.9
C	MoCo-v2	ResNet-50	Aircraft	2.95	41.8	79.9	47.7
C	ReLIC-v2	ResNet-50	Aircraft	2.95	64.8	88.7	26.9
C	SimCLR	ResNet-50	Aircraft	2.95	49.8	87.6	43.2
C	SwAV	ResNet-50	Aircraft	2.95	54.0	83.1	35.0
C	TWIST	ResNet-50	Aircraft	2.95	53.6	85.7	59.9

"linear separability because they maximize dot-product similarity" (Dubois et al. 2024). Furthermore, FT is superior in all self-pretraining cases. Interestingly, for domain shifts with MMD below 1 (Caltech-101 and DTD), LP is mostly superior. FT could in general distort high-quality pretrained features given distribution shifts due to random or zero head initialization which is characterized by a high alignment error and can causes feature distortion (Kumar et al. 2022a). Improving head initialization or a coupled LP-FT approach could therefore also increase FT out-of-distribution performance (Kumar et al. 2022a). At the same time, it should be noted that only CL results are available for Caltech-101 and DTD, so they are therefore not necessarily generalizable to generative models. For higher

MMD values, there is only one outlier with a negative gain (MoCo-v2 with ResNet-50 and CIFAR-100), otherwise FT is superior. This is in accordance with Ericsson et al. (2022b) and Zhang et al. (2023c), who state that FT is better at bridging domain as well as task gaps. FT could be more beneficial if pretext task or data are not perfectly aligned with the downstream objective as it enables task-specific adaptation, domain mismatch correction and performance enhancements (Chen et al. 2020c; Chi et al. 2021; Devlin et al. 2018; Ericsson et al. 2022b; Hu et al. 2020; Pathak et al. 2016; Tao et al. 2023). Given e.g. the significant difference between ImageNet and remote sensing datasets, the fine-tuned version is always superior (Gao et al. 2022). According to He et al. (2023), for BYOL, Rotation and Models Genesis, LP is not beneficial given large data differences (brain MR and cardiac CT). Especially MIM-based approaches seem to profit from FT given large-domain shifts (Gao et al. 2022; Tao et al. 2023). Sorted by ascending domain discrepancy, ImageNet has an average gain of 14.6%, Caltech-101 −1.1%, DTD −3.2%, CUB 30.2%, Flowers 2.8%, UCM 7.2%, CIFAR-10 6.6%, CIFAR-100 18.2%, Aircraft 40.8%. Probably a certain threshold of discrepancy must be reached for FT to be superior to LP, but MMD and average gain do not grow monotonically. Still, the results may be biased insofar as the results of generative models, which tend to have a higher gain, are not available for all datasets.

Domain-shifts both impact the accuracy and the ranking of the best transfer method (Marks et al. 2025). ID transfer methods have a higher Spearman rank correlation compared to their OOD counterparts (Marks et al. 2025). Both in ID and OOD settings, kNN and LP show a higher correlation with each other than with FT (Marks et al. 2025). ID-LP is a good predictor for OOD-LP performance, while ID-FT is only weakly correlated with OOD-FT (Marks et al. 2025). Few-shot FT (10%) is more helpful to estimate the performance of OOD-FT (Marks et al. 2025). Marks et al. (2025) furthermore distinguish two types of domain shifts: style and categorical shifts (with coarse-grained or fine-grained features). The ID performance is a better predictor for the OD performance given categorical shifts than style shifts.

13 Conclusion

Design choices influence performance, generalization and robustness of SSL models substantially. The high level of interdependence of design choices makes their investigation complex and the analysis of the underlying causes of design choice sensitivity essential. Despite their importance, this survey is the first systematic review which contributes to a better understanding of design choices for SSL in CV. This conclusion summarizes the findings of the preceding chapters into four principles of successful SSL design (illustrated in Fig. 12): (1) domain alignment and data focus, (2) a high sample count, meaning that the model is trained with a sufficiently high number of images or variations during pretraining, (3) large model pretraining ensuring an appropriate model capacity, and (4) a pretext representations that has learned pretext invariance aligned with the downstream task. Since the datasets for a specific use case are usually predetermined, design choices should be made from a data perspective. All learning paradigms perform well provided domain-aligned pretraining and domain knowledge is a "key ingredient for successful" SSL (Kang et al. 2023). Hence, domain alignment and data focus are important: they can be implemented via adapted masking, tailored augmentation strategies and samples with a higher relevance

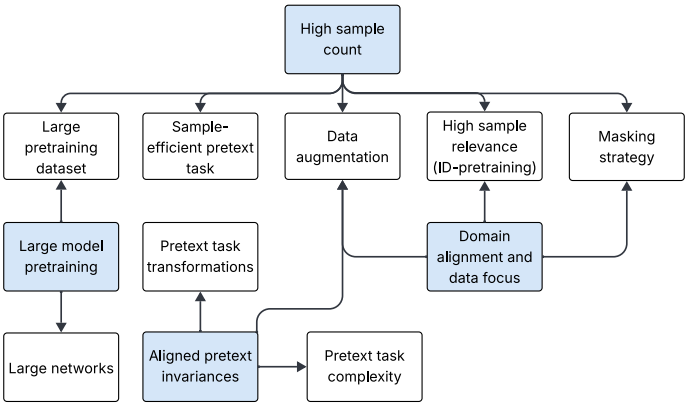

Fig. 12 Important principles of successful design choices for SSL are large model pretraining, high sample count, aligned pretext invariances and domain alignment as well as data focus. Although they cannot always be fulfilled, (e.g., domain-specific and large datasets are not always available), they contribute to improving performance

for the downstream task (ID-pretraining). Increasing the sample count is also important to improve SSL performance (Moutakanni et al. 2024), e.g. via sample-efficient pretext method such as MIM (El-Nouby et al. 2024; Moutakanni et al. 2024). While network size seems to be more important than dataset size for SSL performance, ideally both should be increased accordingly, allowing large model pretraining. Lastly, the pretext representation's equivariances and invariances have to be aligned with the downstream objective and the downstream dataset properties to ensure that it is useful for the downstream task (Ericsson et al. 2022b). This alignment can be ensured via the applied pretext task transformations, the pretext task complexity, and the data augmentation strategy.

13.1 Dependencies between design choices

Understanding how choices interact can guide the development of more reliable and efficient SSL frameworks. Figure 13 presents important dependencies between design choices. The selected downstream data influences the pretraining data choice; both influence the size ratio between them, the data quality and the domain gap. Those three elements in turn influence whether OOD-, ID- or self-pretraining are the best option (see Fig. 14). If LP or FT should be applied for KT is influenced by the dataset sizes, the domain gap and the learning paradigm. The backbone architecture should be in accordance with the learning paradigm and the network capacity, while network capacity should take backbone architecture, computational resources, task complexity and dataset sizes into consideration. The patch size depends on the backbone architecture and on the masking ratio, which in turn is influenced by the masking strategy. The learning paradigm should be chosen according to the data characteristics and the training objective and align with the pretext task design. Training objectives determine the task complexity, which in turn influences pretext task design, pretext complexity and the application of multiple pretext tasks. Each learning paradigm requires different design choices, some of these are also interdependent just like encoder and projector architecture as well as augmentation strategy and augmentation strength.

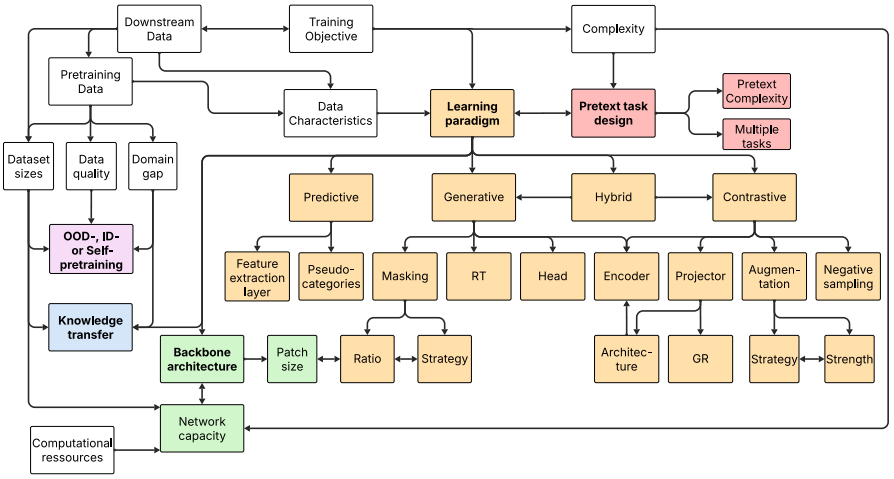

Fig. 13 Important dependencies between design choices that should be considered. White boxes signify factors that are normally fixed given a specific project. Orange boxes are design choices related to the learning paradigm, red boxes to the pretext task design, purple to the datasets, blue to KT and green to the backbone architecture

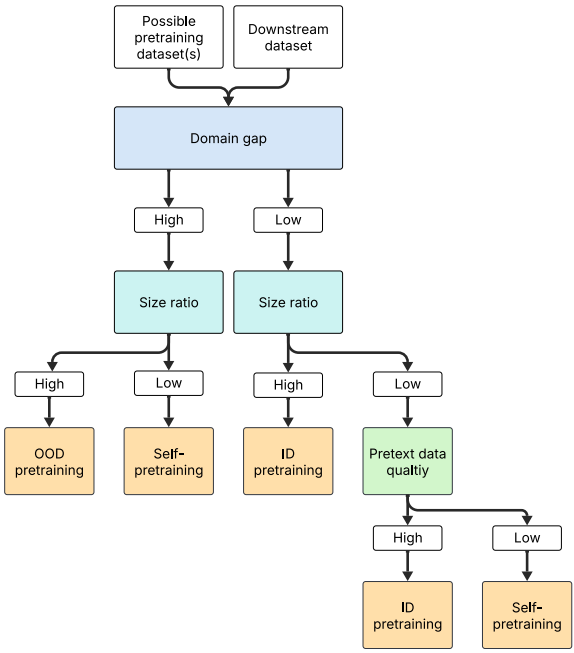

Fig. 14 Decision chart whether OOD, ID or self-pretraining are more recommendable taking the domain gap, size ratio between pretraining and downstream dataset and data quality of the pretext data into consideration

13.2 Practical considerations

As discussed in Sect. 13.1, the complex interactions among design choices make it impossible to define universal strategies that always lead to optimal SSL performance. Nevertheless, the following practical recommendations can be derived from this work:

1. **Higher network capacities**: Increasing the network capacity mostly improves the performance and is therefore advisable, especially in combination with challenging pretext tasks and large pretext datasets (Chen et al. 2020c; Ericsson et al. 2022b; Devlin et al. 2018; Goyal et al. 2019; Jing and Tian 2020). The average improvement on ImageNet-1K for the available scenarios in Sect. 7 is e.g. 6.7% for LP and 2.0% for FT when increasing from ViT-S to ViT-B and 5.1% for LP when increasing from ResNet-50 to ResNet-50 (2x). Smaller architectures and LP seem to profit more from upscaling than larger architectures and FT. It is therefore recommendable using at least ViT-B or ViT-L for LP. The extent to which different SSL frameworks profit from bigger networks varies significantly, some frameworks are considerably more robust regarding small networks.
2. **Larger pretraining datasets:** Training with bigger pretraining datasets seems to be mostly beneficial. In the majority (86%) of the analyzed cases in Table 11, using a larger pretraining datasets is either positive or at least neutral, resulting in an average improvement of 1.5% when comparing the largest and the smallest available dataset on diverse downstream datasets and tasks. Increasing model capacity and dataset size should be considered jointly as one factor can bottleneck the other (Filiot et al. 2023; Xie et al. 2023d) and lager pretraining datasets seem to be most important for large models (Baevski et al. 2023; Goyal et al. 2019; Oquab et al. 2024; Singh et al. 2023). Still, bigger networks seem to be more important than bigger pretraining datasets.
3. **ID pretraining:** ID pretraining seems to be more important than the size of the pretraining dataset. According to Table 12, the average improvement of ID-pretraining across multiple downstream tasks and datasets is 7.7%. ID pretraining is especially recommendable when the available ODD dataset is smaller or not significantly bigger than the ID dataset (ratio 1.1 or below). Given the good performance of ID-pretraining, suitable pretraining datasets could be more important than optimizing the pretext task. Still, there seems to be no linear relation between OOD/ID size ratio and performance gain and ID-pretraining can also be disadvantageous.
4. **Self-pretraining:** Self-pretraining can be particularly useful if only a small OOD dataset is available. According to Table 13, self-pretraining is beneficial or only slightly negative if the self-pretraining dataset is at most about 20 times smaller than the pretraining dataset (on average 21.6% improvement). Self-pretraining is disadvantageous if both the downstream dataset and the domain gap are small.
5. **Pretraining dataset choice:** In practice, the downstream dataset is usually fixed and (depending on the target domain) there is only a limited number of possible pretraining datasets. Therefore, the choice whether to apply OOD, ID or self-pretraining should be based on the available datasets and their characteristics. Figure 14 offers a decision chart based on the research results. The domain gap can e.g. be measured using the Maximum Mean Discrepancy (MMD) between the pretraining ($\mathcal{S}_{\text{pre}}$) and the downstream dataset ($\mathcal{S}_{\text{down}}$) (Chen et al. 2022b):

$$\Delta_{\text{domain}} = \text{MMD}(\mathcal{S}_{\text{pre}}, \mathcal{S}_{\text{down}})$$

The size ratio is the quotient of the number of pretraining and downstream samples:

$$r_{\text{size}} = \frac{N_{\text{pre}}}{N_{\text{down}}}$$

If the domain gap is high but the pretraining dataset is considerably bigger than the downstream dataset, OOD pretraining is probably the best option. If the size ratio is low, then self-pretraining is probably better. Given a low domain gap and a high size ratio, ID pretraining is recommendable. If the size ratio is low, the pretext data quality should be considered, e.g. taking the balance and diversity into consideration, as these factors matter for generalization given low domain gaps (Al Kader Hammoud et al. 2025; Li et al. 2021d). Given a high quality, ID pretraining is probably superior, otherwise self-pretraining is advisable.

6. **FT for self-pretraining:** All analyzed models have a considerably higher FT than LP performance (FT is on average 14.9% better compared to LP) given self-pretraining on ImageNet (see Table 17 and 18). MIM models benefit even more than CL. Given small domain shifts ($0 < \text{MMD} < 1$), LP seems to be advantageous, while for larger domain shifts, FT also seems to be the better option. ID-LP is predictive for OOD-LP performance, whereas ID-FT exhibits only a weak correlation with OOD-FT (Marks et al. 2025).
7. **No superior paradigm:** No SSL paradigm seems to be superior across domains, architectures, datasets and downstream tasks. In practice, selecting the optimal learning paradigm and pretext task should therefor be less relevant than curating extensive and domain-aligned pretraining datasets (Kang et al. 2023). All paradigms require specific design choices.
8. **CL requires optimized augmentations and GR:** Optimizing the augmentation strategy is essential for CL and can generate more significant improvements (ca. 1-5%) than hyperparameter optimization or algorithmic enhancements (Morningstar et al. 2024; Wagner et al. 2022; Zhang et al. 2021e). Discarding the projector after pretraining is crucial to ensure that the more generalizable encoder representation is beneficial for the downstream task (Balestriero et al. 2023; Cosentino et al. 2022a; Gupta et al. 2022; Ma et al. 2023; Mialon et al. 2024; Przewikeźlikowski et al. 2024; Shwartz Ziv and LeCun 2024; Xue et al. 2024). Projectors mitigate eventual biases due to pretraining or ill optimization. The optimal layers to discard during GR depend on the downstream task, dataset, training configuration and pretext overfitting (Bordes et al. 2023a). Increasing the dimensionality of the last encoder layer increases the performance (Assran et al. 2022a; Bordes et al. 2023c; Dubois et al. 2024).
9. **MIM benefit from ViT and moderate masking ratios:** MIM models should preferably use large-capacity ViT backbones (El-Nouby et al. 2021; Li et al. 2024a; Konstantakos et al. 2025; Wei et al. 2022b; Xie et al. 2023d). They require masking strategies to learn higher-level semantic features. Very high masking ratios ($\geq 90\%$) do not seem to be beneficial, probably they hinder higher-level learning (Kong et al. 2023; Li et al. 2024b; Xie et al. 2022b). Very small masking ratios ($\leq 15\%$) seem to be only beneficial for small networks (e.g. ViT-S) or high patch ratios. Most MIM models seem to benefit

from moderate masking ratios. Similar to projectors in CL, MIM profits from discarding the head after pretraining.

10. **MIM works best with FT:** While FT is most suitable for MIM, self-pretraining and large domain gaps, LP can be beneficial for CL and small domain gaps.
11. **Predictive SSL requires distinctive pseudo-categories:** The distinctiveness of pseudo-categories is important for predictive SSL and should be considered for pretext design.During pretraining, the representation quality can decrease in deeper layers and becomes too adapted to the pretext task, so features should rather be extracted in intermediate layers to improve downstream transferability (Caron et al. 2018; Gidaris et al. 2018; Kim et al. 2018; Kolesnikov et al. 2019; Mundhenk et al. 2018; Noroozi and Favaro 2016; Zhang et al. 2023c; Zhuang et al. 2019b). Multiple pretext task can be beneficial if these are well aligned.

13.3 Limitations

Similar to hyperparameters, the gains from optimizing one design choice are often nonlinear and influenced by other design choices, which makes it difficult to transition from one design setup to another (Moutakanni et al. 2024). Determining which design choice has the greatest impact is challenging, as one single choice often is not isolated. Consequently, negative effects of one suboptimal design choice could be mitigated by another. Most publications vary several design choices and hyperparameters, making it challenging to draw clear conclusions. Even if such conclusions are possible for one specific case, they cannot always be generalized. Furthermore, it is difficult to determine which design choice exerts the most significant influence due to the difficulty in comparing the varying degrees of change in different design choices. For example, how to you compare an increase from ViT-B to ViT-L with an increase in dataset size from ImageNet-1K to ImageNet-21K when in the first case the relevant unit is “trainable parameters” and in the second case it is “data samples”? As research on SSL often focuses on few object-centric RGB datasets, such as ImageNet, findings can not always be generalized to other datasets (Cole et al. 2022) and semantic information might not be preserved during SSL given images from other domains (Huang et al. 2023). Developing wider benchmarks to study SSL performance across multiple tasks and domains is therefore important to avoid designing and optimizing SSL only for few object-centered datasets (Ericsson et al. 2022b). Varying design choices and hyperparameter settings across publications may also explain why research results sometimes appear contradictory.

13.4 Future work

This work lays the foundation for a better understanding of design choices for SSL, but further research is required. The following aspects in particular should be investigated:

1. **Isolated and interacting effects**: While the systematic examination of an isolated design choice can be of great interest, the interconnections should always be taken into account to better understand sensitivity and dependencies. Related design choices across paradigms should be jointly considered, as often phenomena are perceived as distinct when, in fact, they are similar or even identical. This is often due to different

terminologies or research contexts. For instance, the concepts of extraction layers and GR are regarded as distinct, despite sharing similarities. It would be worthwhile to study such design choices together to create a more holistic understanding.

2. **Design choices and robustness**: While this work mainly focuses on the impact of design choices on performance, future work should analyze how design choices influence robusteness, i.e. with respect to input perturbations, distribution shifts, and downstream task variability.
3. **Disentangling backbone choice and pretext objective**: More research is required to determine the extent to which the differences between contrastive and generative SSL are due to the fact that the former is mostly used with CNN backbones and the latter with ViT backbones, and to what extent they are actually caused by the different pretext tasks.
4. **Design choices of hybrid SSL**: Hybrid SSL could in general overcome the limitations of the individual paradigms (Deepa et al. 2025; Shi et al. 2023b) and will probably continue to increase importance. Consequently, a focus on design choices for hybrid SSL and how combining generative and discriminative elements creates synergies would be interesting.
5. **Transparent reporting and reproducibility**: It is particularly important that design choices are made transparent in all publications in order to create a basis for comparison and reproducibility.

Ultimately, deeper insight into design choices will be essential not only for improving performance, but also for achieving robustness and reliability across diverse downstream tasks and data distributions.

Author contributions This paper was written exclusively by L.W.

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was received to assist with the preparation of this manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The author has no conflict of interest to declare that are relevant to the content of this article.

References

Abdulrazzaq MM, Ramaha NT, Hameed AA et al (2024) Consequential advancements of self-supervised learning (ssl) in deep learning contexts. Mathematics 12(5):758. https://doi.org/10.3390/math12050758

Aberdam A, Litman R, Tsiper S, et al (2021) Sequence-to-sequence contrastive learning for text recognition. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 15297–15307, https://doi.org/10.1109/CVPR46437.2021.01505

Abnar S, Dehghani M, Neyshabur B, et al (2021) Exploring the limits of large scale pre-training. arXiv:2110.02095

Aerts H, Velazquez ER, Leijenaar R et al (2015) Data from nsclc-radiomics. Cancer Imaging Arch. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI

Ahn WJ, Yang GY, Choi HD, et al (2022) Improving vision transformer with multi-task training. In: 2022 22nd International conference on control, automation and systems (ICCAS), pp 1963–1965, https://doi.org/10.23919/ICCAS55662.2022.10003833

Al Kader Hammoud HA, Das T, Pizzati F, et al (2025) On pretraining data diversity for self-supervised learning. In: Computer vision – ECCV 2024. Springer Nature Switzerland, pp 54–71, https://doi.org/10.1007/978-3-031-72992-8_4

Al Mamun A, Ahmedt-Aristizabal D, Zhang M et al (2024) Plant disease detection using self-supervised learning: a systematic review. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3475819

Albelwi S (2022) Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging. Entropy. https://doi.org/10.3390/e24040551

Alkin B, Miklautz L, Hochreiter S, et al (2025) Mim-refiner: A contrastive learning boost from intermediate pre-trained representations. https://doi.org/10.48550/arXiv.2402.10093

Almalki A, Latecki LJ (2023) Self-supervised learning with masked image modeling for teeth numbering, detection of dental restorations, and instance segmentation in dental panoramic radiographs. In: 2023 IEEE/CVF winter conference on applications of computer vision (WACV), pp 5583–5592, https://doi.org/10.1109/WACV56688.2023.00555

Amrani E, Karlinsky L, Bronstein A (2022) Self-supervised classification network. In: Computer vision – ECCV 2022. Springer Nature Switzerland, Cham, pp 116–132, https://doi.org/10.1007/978-3-031-19821-2_7

Anton J, Castelli L, Chan MF et al (2022) How well do self-supervised models transfer to medical imaging? J Imaging. https://doi.org/10.3390/jimaging8120320

Appalaraju S, Zhu Y, Xie Y, et al (2020) Towards good practices in self-supervised representation learning. https://doi.org/10.48550/arXiv.2012.00868

Arboretti R, Ceccato R, Pegoraro L et al (2022a) Design choice and machine learning model performances. Qual Reliab Eng Int 38(7):3357–3378. https://doi.org/10.1002/qre.3123

Arboretti R, Ceccato R, Pegoraro L et al (2022b) Design of experiments and machine learning for product innovation: a systematic literature review. Qual Reliab Eng Int 38(2):1131–1156. https://doi.org/10.1002/qre.3025

Aresta G, Araújo T, Kwok S et al (2019) Bach: Grand challenge on breast cancer histology images. Med Image Anal 56:122–139. https://doi.org/10.1016/j.media.2019.05.010

Arora S, Khandeparkar H, Khodak M, et al (2019) A theoretical analysis of contrastive unsupervised representation learning. arXiv preprint https://doi.org/10.48550/arXiv.1902.09229

Assran M, Balestriero R, Duval Q, et al (2022a) The hidden uniform cluster prior in self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2210.07277

Assran M, Caron M, Misra I, et al (2022b) Masked siamese networks for label-efficient learning. In: European conference on computer vision. Springer Nature Switzerland, pp 456–473, https://doi.org/10.1007/978-3-031-19821-2_26

Assran M, Duval Q, Misra I, et al (2023) Self-supervised learning from images with a joint-embedding predictive architecture. In: 2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 15619–15629, https://doi.org/10.1109/CVPR52729.2023.01499

Atito S, Awais M, Kittler J (2022) Sit: Self-supervised vision transformer. arXiv preprint https://doi.org/10.48550/arXiv.2104.03602

Auh J, Cho C, St K (2024) Improved contrastive learning model via identification of false-negatives in self-supervised learning. ETRI J 46(6):1020–1029. https://doi.org/10.4218/etrij.2023-0285

Awais M, Naseer M, Khan S et al (2025) Foundation models defining a new era in vision: a survey and outlook. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2024.3506283

Ayromlou S, Khazaie VR, Forghani F et al (2025) Can generative models improve self-supervised representation learning? Proc AAAI Conf Artif Intell 39(2):1800–1808. https://doi.org/10.1609/aaai.v39i2.32174

Azizi S, Mustafa B, Ryan F, et al (2021) Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3478–3488, https://doi.org/10.1109/ICCV48922.2021.00346

Azizi S, Culp L, Freyberg J et al (2023) Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nat Biomed Eng 7(6):756–779. https://doi.org/10.1038/s41551-023-01049-7

Babaev D, Ovsov N, Kireev I, et al (2022) Coles: Contrastive learning for event sequences with self-supervision. In: Proceedings of the 2022 international conference on management of data, pp 1190–1199, https://doi.org/10.1145/3514221.3526129

Bachman P, Hjelm RD, Buchwalter W (2019) Learning representations by maximizing mutual information across views. arXiv preprint https://doi.org/10.48550/arXiv.1906.00910

Bachmann R, Mizrahi D, Atanov A, et al (2022) Multimae: Multi-modal multi-task masked autoencoders. In: European conference on computer vision, Springer, pp 348–367, https://doi.org/10.1007/978-3-031-19836-6_20

Baevski A, Hsu WN, Xu Q, et al (2022) Data2vec: A general framework for self-supervised learning in speech, vision and language. In: International conference on machine learning, PMLR, pp 1298–1312, https://proceedings.mlr.press/v162/baevski22a.html

Baevski A, Babu A, Hsu WN, et al (2023) Efficient self-supervised learning with contextualized target representations for vision, speech and language. In: International conference on machine learning, PMLR, pp 1416–1429, https://proceedings.mlr.press/v202/baevski23a.html

Bai Y, Yang Y, Zhang W, et al (2022) Directional self-supervised learning for heavy image augmentations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16692–16701

Bai Y, Geng X, Mangalam K, et al (2024) Sequential modeling enables scalable learning for large vision models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 22861–22872

Bai L, Zhang X, Qin W, et al (2025) From orbit to ground: A comprehensive review of multimodal self-supervised learning for remote sensing. Authorea Preprints https://doi.org/10.36227/techrxiv.174137752.20965470/v1

Bakr S, Gevaert O, Echegaray S et al (2018) A radiogenomic dataset of non-small cell lung cancer. Sci Data 5(1):1–9. https://doi.org/10.1038/sdata.2018.202

Balestriero R, Ibrahim M, Sobal V, et al (2023) A cookbook of self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2304.12210

Balmaseda V, Wang B, Lin CL, et al (2025) Discovering global false negatives on the fly for self-supervised contrastive learning. arXiv preprint https://doi.org/10.48550/arXiv.2502.20612

Bandara WGC, De Melo CM, Patel VM (2023) Guarding barlow twins against overfitting with mixed samples. arXiv preprint https://doi.org/10.48550/arXiv.2312.02151

Bao H, Dong L, Piao S, et al (2021) Beit: Bert pre-training of image transformers. arXiv preprint https://doi.org/10.48550/arXiv.2106.08254

Baraldi L, Amoroso R, Cornia M, et al (2023) Learning to mask and permute visual tokens for vision transformer pre-training. arXiv preprint https://doi.org/10.48550/arXiv.2306.07346

Bardes A, Ponce J, LeCun Y (2021) Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2105.04906

Bardes A, Ponce J, LeCun Y (2022) Vicregl: Self-supervised learning of local visual features. arXiv preprint https://doi.org/10.48550/arXiv.2210.01571

Baykal G, Unal G (2020) Deshufflegan: A self-supervised gan to improve structure learning. In: 2020 IEEE international conference on image processing (ICIP), IEEE, pp 708–712, https://doi.org/10.1109/ICIP40778.2020.9190774

Baykal G, Ozcelik F, Unal G (2022) Exploring deshufflegans in self-supervised generative adversarial networks. Pattern Recogn 122:108244. https://doi.org/10.1016/j.patcog.2021.108244

Beck MA, Liu CY, Bidinosti CP, et al (2021) Presenting an extensive lab-and field-image dataset of crops and weeds for computer vision tasks in agriculture. arXiv preprint https://doi.org/10.48550/arXiv.2108.05789

Bendidi I, Bardes A, Cohen E, et al (2023) No free lunch in self supervised representation learning. arXiv preprint https://doi.org/10.48550/arXiv.2304.11718

Bendidi I, Bardes A, Cohen E et al (2024) Exploring self-supervised learning biases for microscopy image representation. Biol Imaging 4:e12. https://doi.org/10.1017/S2633903X2400014X

Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Benigmim Y, Roy S, Essid S, et al (2024) Collaborating foundation models for domain generalized semantic segmentation. In: 2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR), IEEE, pp 3108–3119, https://doi.org/10.1109/CVPR52733.2024.00300

Berg P, Pham MT, Courty N (2022) Self-supervised learning for scene classification in remote sensing: current state of the art and perspectives. Remote Sens 14(16):3995. https://doi.org/10.3390/rs14163995
Bernard O, Lalande A, Zotti C et al (2018) Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging 37(11):2514–2525. https://doi.org/10.1109/TMI.2018.2837502
Betker J, Goh G, Jing L et al (2023) Improving image generation with better captions. Comput Sci 2(3):8
Bien N, Rajpurkar P, Ball RL et al (2018) Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. PLoS Med 15(11):e1002699. https://doi.org/10.1371/journal.pmed.1002699
Bommasani R (2023) On the opportunities and risks of foundation models. arXiv preprint https://doi.org/10.48550/arXiv.2108.07258
Bordes F, Balestriero R, Garrido Q et al (2023a) Guillotine regularization: why removing layers is needed to improve generalization in self-supervised learning. Trans Mach Learn Res. https://doi.org/10.48550/arXiv.2206.13378
Bordes F, Balestriero R, Vincent P (2023b) Towards democratizing joint-embedding self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2303.01986
Bordes F, Lavoie S, Balestriero R, et al (2023c) A surprisingly simple technique to control the pretraining bias for better transfer: expand or narrow your representation. arXiv preprint https://doi.org/10.48550/arXiv.2304.05369
Bossard L, Guillaumin M, Van Gool L (2014) Food-101–mining discriminative components with random forests. In: Computer vision–ECCV 2014. Springer International Publishing, pp 446–461, https://doi.org/10.1007/978-3-319-10599-4_29
Boulahia SY, Amamra A, Madi MR et al (2021) Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Mach Vis Appl 32(6):121. https://doi.org/10.1007/s00138-021-01249-8
Brown TB, Mann B, Ryder N, et al (2020) Language models are few-shot learners. arXiv preprint https://doi.org/10.48550/arXiv.2005.14165
Cabannes V, Kiani B, Balestriero R, et al (2023) The ssl interplay: augmentations, inductive bias, and generalization. In: International conference on machine learning, PMLR, pp 3252–3298
Çağatan ÖV, Tal ÖF, Gursoy ME (2025) Adversarial robustness of discriminative self-supervised learning in vision. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2313–2324
Campanella G, Vanderbilt C, Fuchs T (2024) Computational pathology at health system scale–self-supervised foundation models from billions of images. In: AAAI 2024 spring symposium on clinical foundation models
Campanella G, Chen S, Singh M et al (2025) A clinical benchmark of public self-supervised pathology foundation models. Nat Commun 16(1):3640. https://doi.org/10.1038/s41467-025-58796-1
Cao YH, Wu J (2021) Rethinking self-supervised learning: small is beautiful. arXiv preprint https://doi.org/10.48550/arXiv.2103.13559
Cao S, Xu P, Clifton DA (2022) How to understand masked autoencoders. arXiv preprint https://doi.org/10.48550/arXiv.2202.03670
Caron M, Bojanowski P, Joulin A, et al (2018) Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). Springer International Publishing, pp 139–156, https://doi.org/10.1007/978-3-030-01264-9_9
Caron M, Misra I, Mairal J, et al (2021a) Unsupervised learning of visual features by contrasting cluster assignments. https://doi.org/10.48550/arXiv.2006.09882
Caron M, Touvron H, Misra I, et al (2021b) Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9650–9660, https://doi.org/10.1109/ICCV48922.2021.00951
Chander B, John C, Warrier L et al (2025) Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness. ACM Comput Surv 57(6):1–49. https://doi.org/10.1145/3675392
Chang H, Zhang H, Jiang L, et al (2022) Maskgit: Masked generative image transformer. arXiv preview https://doi.org/10.48550/arXiv.2202.04200
Chen X, He K (2021) Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15750–15758, https://doi.org/10.1109/CVPR46437.2021.01549
Chen X, Wang S, Fu B, et al (2019a) Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. Adv Neural Inform Process Syst 32
Chen T, Zhai X, Ritter M, et al (2019b) Self-supervised gans via auxiliary rotation loss. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12146–12155, https://doi.org/10.1109/CVPR.2019.01243

Chen X, Fan H, Girshick R, et al (2020a) Improved baselines with momentum contrastive learning. arXiv preprint https://doi.org/10.48550/arXiv.2003.04297
Chen T, Kornblith S, Norouzi M, et al (2020b) A simple framework for contrastive learning of visual representations. https://doi.org/10.48550/arXiv.2002.05709
Chen T, Kornblith S, Swersky K, et al (2020c) Big self-supervised models are strong semi-supervised learners. https://doi.org/10.48550/arXiv.2006.10029
Chen M, Radford A, Child R, et al (2020d) Generative pretraining from pixels. In: International conference on machine learning, PMLR, pp 1691–1703
Chen T, Frankle J, Chang S, et al (2021a) The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16306–16316
Chen T, Luo C, Li L (2021b) Intriguing properties of contrastive losses. Adv Neural Inf Process Syst 34:11834–11845
Chen X, Xie S, He K (2021c) An empirical study of training self-supervised vision transformers. In: 2021 IEEE/CVF international conference on computer vision (ICCV), pp 9620–9629, https://doi.org/10.1109/ICCV48922.2021.00950
Chen Z, Du Y, Hu J, et al (2022a) Multi-modal masked autoencoders for medical vision-and-language pre-training. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 679–689, https://doi.org/10.1007/978-3-031-16443-9_65
Chen Y, Li J, Ding C, et al (2022b) Rethinking two consensuses of the transferability in deep learning. arXiv preprint https://doi.org/10.48550/arXiv.2212.00399
Chen Y, Liu Y, Jiang D, et al (2022c) Sdae: Self-distillated masked autoencoder. In: European conference on computer vision, Springer, pp 108–124, https://doi.org/10.1007/978-3-031-20056-4_7
Chen Y, Mancini M, Zhu X et al (2022d) Semi-supervised and unsupervised deep visual learning: a survey. IEEE Trans Pattern Anal Mach Intell 46(3):1327–1347. https://doi.org/10.1109/TPAMI.2022.3201576
Chen C, Zhang J, Xu Y et al (2022e) Why do we need large batchsizes in contrastive learning? A gradient-bias perspective. Adv Neural Inf Process Syst 35:33860–33875
Chen K, Liu Z, Hong L, et al (2023a) Mixed autoencoder for self-supervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 22742–22751, https://doi.org/10.1109/CVPR52729.2023.02178
Chen H, Zhang W, Wang Y, et al (2023b) Improving masked autoencoders by learning where to mask. In: Chinese conference on pattern recognition and computer vision (PRCV), Springer, pp 377–390, https://doi.org/10.1007/978-981-99-8543-2_31
Chen C, Zhong A, Wu D, et al (2023c) Contrastive masked image-text modeling for medical visual representation learning. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 493–503, https://doi.org/10.1007/978-3-031-43904-9_48
Chen X, Ding M, Wang X et al (2024a) Context autoencoder for self-supervised representation learning. Int J Comput Vis 132(1):208–223. https://doi.org/10.1007/s11263-023-01852-4
Chen Z, Hu B, Chen Z, et al (2024b) Progress and thinking on self-supervised learning methods in computer vision: a review. IEEE Sens J
Chi Z, Huang H, Liu L, et al (2021) Cross-lingual language model meta-pretraining. arXiv preprint https://doi.org/10.48550/arXiv.2109.11129
Chin ZY, Jiang CM, Huang CC, et al (2024) Masking improves contrastive self-supervised learning for convnets, and saliency tells you where. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2761–2770, https://doi.org/10.1109/WACV57701.2024.00274
Choi K (2024) Self-supervised learning for ct image denoising and reconstruction: a review. Biomed Eng Lett 14(6):1207–1220. https://doi.org/10.1007/s13534-024-00424-w
Choi H, Park H, Yi KM, et al (2024) Salience-based adaptive masking: revisiting token dynamics for enhanced pre-training. arXiv preprint https://doi.org/10.48550/arXiv.2404.08327
Chowdhury A, Rosenthal J, Waring J, et al (2021) Applying self-supervised learning to medicine: review of the state of the art and medical implementations. In: Informatics, MDPI, p 59, https://doi.org/10.3390/informatics8030059
Cimpoi M, Maji S, Kokkinos I, et al (2014) Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3606–3613, https://doi.org/10.1109/CVPR.2014.461
Ciocarlan A, Lefebvre S, Hégarat-Mascle SL, et al (2024) Self-supervised learning for real-world object detection: a survey. arXiv preprint
Coates A, Ng A, Lee H (2011) An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, pp 215–223

Cole E, Deneu B, Lorieul T, et al (2020) The geolifeclef 2020 dataset. arXiv preprint https://doi.org/10.48550/arXiv.2004.04192
Cole E, Yang X, Wilber K, et al (2022) When does contrastive visual representation learning work? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14755–14764, https://doi.org/10.1109/CVPR52688.2022.01434
Cordts M, Omran M, Ramos S, et al (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3213–3223, https://doi.org/10.1109/CVPR.2016.350
Cosentino R, Sengupta A, Avestimehr S, et al (2022a) Toward a geometrical understanding of self-supervised contrastive learning. arXiv preprint https://doi.org/10.48550/arXiv.2205.06926
Cosentino R, Shekkizhar S, Soltanolkotabi M, et al (2022b) The geometry of self-supervised learning models and its impact on transfer learning. arXiv preprint https://doi.org/10.48550/arXiv.2209.08622
Cubuk ED, Zoph B, Mane D, et al (2018) Autoaugment: Learning augmentation policies from data. arXiv preprint https://doi.org/10.48550/arXiv.1805.09501
Dangovski R, Jing L, Loh C, et al (2021) Equivariant contrastive learning. arXiv preprint https://doi.org/10.48550/arXiv.2111.00899
Deepa S, Sheetal AA et al (2025) Enhancing image classification performance through hybrid self-supervised learning strategies. SSRG Int J Electron Commun Eng. https://doi.org/10.48550/arXiv.2206.02353
Deldari S, Xue H, Saeed A, et al (2022) Beyond just vision: a review on self-supervised representation learning on multimodal and temporal data. arXiv preprint https://doi.org/10.48550/arXiv.2206.02353
Deng L (2012) The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process Mag 29(6):141–142. https://doi.org/10.1109/MSP.2012.2211477
Deng J, Dong W, Socher R, et al (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255, https://doi.org/10.1109/CVPR.2009.5206848
Devlin J, Chang MW, Lee K, et al (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint https://doi.org/10.48550/arXiv.1810.04805
Dhariwal P, Nichol A (2021) Diffusion models beat GANs on image synthesis. https://doi.org/10.48550/arXiv.2105.05233
Dimitrovski I, Kitanovski I, Simidjievski N et al (2024) In-domain self-supervised learning improves remote sensing image scene classification. IEEE Geosci Remote Sens Lett. https://doi.org/10.1109/LGRS.2024.3352926
Doersch C, Zisserman A (2017) Multi-task self-supervised visual learning. In: Proceedings of the IEEE international conference on computer vision, pp 2051–2060, https://doi.org/10.1109/ICCV.2017.226
Doersch C, Gupta A, Efros AA (2015) Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision, pp 1422–1430, https://doi.org/10.1109/ICCV.2015.167
Donahue J, Simonyan K (2019) Large scale adversarial representation learning. https://doi.org/10.48550/arXiv.1907.02544
Donahue J, Krähenbühl P, Darrell T (2016) Adversarial feature learning. arXiv preprint https://doi.org/10.48550/arXiv.1605.09782
Dong X, Bao J, Zhang T, et al (2022) Bootstrapped masked autoencoders for vision bert pretraining. In: European conference on computer vision, Springer, pp 247–264
Dong X, Bao J, Zhang T, et al (2023a) Peco: Perceptual codebook for bert pre-training of vision transformers. In: Proceedings of the AAAI conference on artificial intelligence, pp 552–560, https://doi.org/10.1609/aaai.v37i1.25130
Dong X, Bao J, Zheng Y, et al (2023b) Maskclip: Masked self-distillation advances contrastive language-image pretraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10995–11005, https://doi.org/10.1109/CVPR52729.2023.01058
Dong H, Long X, Li Y (2024) Synthetic hard negative samples for contrastive learning. Neural Process Lett 56(1):33. https://doi.org/10.1007/s11063-024-11522-2
Dong Q, Zhou Z, Qiu X et al (2025) A survey on self-supervised monocular depth estimation based on deep neural networks. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2025.3552598
Dosovitskiy A, Fischer P, Springenberg JT, et al (2015) Discriminative unsupervised feature learning with exemplar convolutional neural networks. https://doi.org/10.48550/arXiv.1406.6909
Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint https://doi.org/10.48550/arXiv.2010.11929
Dubois Y, Hashimoto T, Ermon S, et al (2022) Improving self-supervised learning by characterizing idealized representations. https://doi.org/10.48550/arXiv.2209.06235
Dubois Y, Hashimoto T, Liang P (2024) Evaluating self-supervised learning via risk decomposition. arXiv preprint https://doi.org/10.48550/arXiv.2302.03068

Duesterwald E, Murthi A, Venkataraman G, et al (2019) Exploring the hyperparameter landscape of adversarial robustness. arXiv preprint https://doi.org/10.48550/arXiv.1905.03837
Dufumier B, Castillo-Navarro J, Tuia D, et al (2025) What to align in multimodal contrastive learning? arXiv preprint https://doi.org/10.48550/arXiv.2409.07402
Du T, Wang Y, Wang Y (2024) On the role of discrete tokenization in visual representation learning. arXiv preprint https://doi.org/10.48550/arXiv.2407.09087
Dwibedi D, Aytar Y, Tompson J, et al (2021) 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, https://doi.org/10.1109/ICCV48922.2021.00945
El-Nouby A, Izacard G, Touvron H, et al (2021) Are large-scale datasets necessary for self-supervised pre-training? arXiv preprint https://doi.org/10.48550/arXiv.2112.10740
El-Nouby A, Klein M, Zhai S, et al (2024) Scalable pre-training of large autoregressive image models. arXiv preprint https://doi.org/10.48550/arXiv.2401.08541
El-Shimy H, Zantout H, Lones MA, et al (2025) Self-supervised learning for pre-training capsule networks: overcoming medical imaging dataset challenges. arXiv preprint https://doi.org/10.48550/arXiv.2502.04748
Ericsson L, Gouk H, Hospedales TM (2021) How well do self-supervised models transfer? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5414–5423, https://doi.org/10.1109/CVPR46437.2021.00537
Ericsson L, Gouk H, Hospedales TM (2022a) Why do self-supervised models transfer? Investigating the impact of invariance on downstream tasks. arXiv preprint https://doi.org/10.48550/arXiv.2111.11398
Ericsson L, Gouk H, Loy CC et al (2022b) Self-supervised representation learning: introduction, advances, and challenges. IEEE Signal Process Mag 39(3):42–62. https://doi.org/10.1109/MSP.2021.3134634
Estepa IG, Sarasúa I, Nagarajan B, et al (2023) All4one: Symbiotic neighbour contrastive learning via self-attention and redundancy reduction. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 16243–16253, https://doi.org/10.1109/ICCV51070.2023.01488
Everett M, Zhong M, Leontidis G (2025) Masked capsule autoencoders. arXiv preprint arXiv:2403.04724
Everingham M, Van Gool L, Williams CKI, et al (2007) The PASCAL visual object classes challenge 2007 (VOC2007) results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
Falcon W, Cho K (2020) A framework for contrastive self-supervised learning and designing a new approach. arXiv preprint https://doi.org/10.48550/arXiv.2009.00104
Fang Y, Dong L, Bao H, et al (2022) Corrupted image modeling for self-supervised visual pre-training. arXiv preprint https://doi.org/10.48550/arXiv.2202.03382
Fang A, Jose AM, Jain A, et al (2023a) Data filtering networks. arXiv preprint https://doi.org/10.48550/arXiv.2309.17425
Fang Y, Wang W, Xie B, et al (2023b) Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19358–19369, https://doi.org/10.1109/CVPR52729.2023.01855
Fedorov A, Geenjaar E, Wu L et al (2024) Self-supervised multimodal learning for group inferences from mri data: discovering disorder-relevant brain regions and multimodal links. Neuroimage 285:120485. https://doi.org/10.1016/j.neuroimage.2023.120485
Fei Z, Fan M, Zhu L, et al (2023) Masked auto-encoders meet generative adversarial networks and beyond. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 24449–24459, https://doi.org/10.1109/CVPR52729.2023.02342
Feng Z, Xu C, Tao D (2019) Self-supervised representation learning by rotation feature decoupling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10364–10374, https://doi.org/10.1109/CVPR.2019.01061
Feng R, Zhou Z, Gotway MB, et al (2020) Parts2whole: self-supervised contrastive learning via reconstruction. In: Domain adaptation and representation transfer, and distributed and collaborative learning. Springer, pp 85–95, https://doi.org/10.1007/978-3-030-60548-3_9
Fernandez J, Wehrstedt L, Shamis L, et al (2025) Hardware scaling trends and diminishing returns in large-scale distributed training. arXiv preprint https://doi.org/10.48550/arXiv.2411.13055
Filiot A, Ghermi R, Olivier A, et al (2023) Scaling self-supervised learning for histopathology with masked image modeling. medRxiv pp 2023–07. https://doi.org/10.1101/2023.07.21.23292757
Gao X, Hu W, Qi GJ (2020) Graphter: Unsupervised learning of graph transformation equivariant representations via auto-encoding node-wise transformations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7163–7172, https://doi.org/10.1109/CVPR42600.2020.00719
Gao Y, Sun X, Liu C (2022) A general self-supervised framework for remote sensing image classification. Remote Sens 14(19):4824. https://doi.org/10.3390/rs14194824
Gao P, Lin Z, Zhang R et al (2024) Mimic before reconstruct: enhancing masked autoencoders with feature mimicking. Int J Comput Vis. https://doi.org/10.1007/s11263-023-01898-4

Garrido Q, Chen Y, Bardes A, et al (2022) On the duality between contrastive and non-contrastive self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2206.02574
Garrido Q, Balestriero R, Najman L, et al (2023a) Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. https://doi.org/10.48550/arXiv.2210.02885
Garrido Q, Najman L, Lecun Y (2023b) Self-supervised learning of split invariant equivariant representations. arXiv preprint https://doi.org/10.48550/arXiv.2302.10283
Ghanbarzadeh A, Soleimani H (2024) In-domain supervised and contrastive self-supervised representation learning for dense prediction problems in remote sensing imageries. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3510779
Giakoumoglou N, Stathaki T, Gkelias A (2025) A review on discriminative self-supervised learning methods in computer vision. arXiv preprint https://doi.org/10.48550/arXiv.2405.04969
Gidaris S, Singh P, Komodakis N (2018) Unsupervised representation learning by predicting image rotations. arXiv preprint https://doi.org/10.48550/arXiv.1803.07728
Girdhar R, El-Nouby A, Liu Z, et al (2023) Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15180–15190
Girish S, Dey D, Joshi N, et al (2022) One network doesn't rule them all: moving beyond handcrafted architectures in self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2203.08130
Goldblum M, Souri H, Ni R et al (2023) Battle of the backbones: a large-scale comparison of pretrained models across computer vision tasks. Adv Neural Inf Process Syst 36:29343–29371
Goodfellow I, Pouget-Abadie J, Mirza M et al (2020) Generative adversarial networks. Commun ACM 63(11):139–144. https://doi.org/10.1145/3422622
Goyal N (2022) A survey on self supervised learning approaches for improving multimodal representation learning. arXiv preprint https://doi.org/10.48550/arXiv.2210.11024
Goyal P, Mahajan D, Gupta A, et al (2019) Scaling and benchmarking self-supervised visual representation learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6391–6400, https://doi.org/10.1109/ICCV.2019.00649
Goyal P, Caron M, Lefaudeux B, et al (2021) Self-supervised pretraining of visual features in the wild. arXiv preprint https://doi.org/10.48550/arXiv.2103.01988
Goyal S, Kumar A, Garg S, et al (2023) Finetune like you pretrain: improved finetuning of zero-shot vision models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19338–19347, https://doi.org/10.1109/CVPR52729.2023.01853
Griffin G, Holub A, Perona P, et al (2007) Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology Pasadena, https://doi.org/10.22002/D1.20087
Grill JB, Strub F, Altché F, et al (2020) Bootstrap your own latent: a new approach to self-supervised learning. https://doi.org/10.48550/arXiv.2006.07733
Gui Y, Ma C, Zhong Y (2023) Unraveling projection heads in contrastive learning: insights from expansion and shrinkage. arXiv preprint https://doi.org/10.48550/arXiv.2306.03335
Gui J, Chen T, Cao Q, et al (2024) A survey of self-supervised learning from multiple perspectives: algorithms, theory, applications and future trends. arXiv preprint https://doi.org/10.48550/arXiv.2301.05712
Gu J, Meng X, Lu G, et al (2022) Wukong: A 100 million large-scale Chinese cross-modal pre-training benchmark. https://doi.org/10.48550/arXiv.2202.06767
Gunel B, Du J, Conneau A, et al (2020) Supervised contrastive learning for pre-trained language model fine-tuning. arXiv preprint https://doi.org/10.48550/arXiv.2011.01403
Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: International conference on machine learning, PMLR, pp 1321–1330
Guo W, Huang H, Kong X, et al (2019) Learning disentangled representation for cross-modal retrieval with deep mutual information estimation. In: Proceedings of the 27th ACM international conference on multimedia, pp 1712–1720, https://doi.org/10.1145/3343031.3351053
Guo J, Han K, Wu H, et al (2022) Fastmim: Expediting masked image modeling pre-training for vision. arXiv preprint https://doi.org/10.48550/arXiv.2212.06593
Gupta S (2025) Structuring representation geometry in self-supervised learning. PhD thesis, Massachusetts Institute of Technology
Gupta K, Ajanthan T, Hengel Avd, et al (2022) Understanding and improving the role of projection head in self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2212.11491
Gutmann M, Hyvärinen A (2010) Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, pp 297–304
Haja A, Van Der Woude B, Schomaker L (2023) Organoids segmentation using self-supervised learning: How complex should the pretext task be? In: Proceedings of the 2023 10th international conference on biomedical and bioinformatics engineering, pp 17–27, https://doi.org/10.1145/3637732.3637772

Han X, Zhang Z, Ding N et al (2021) Pre-trained models: past, present and future. AI Open 2:225–250. https://doi.org/10.1016/j.aiopen.2021.08.002
HaoChen JZ, Wei C, Gaidon A, et al (2022a) Provable guarantees for self-supervised deep learning with spectral contrastive loss. arXiv preprint https://doi.org/10.48550/arXiv.2106.04156
HaoChen JZ, Wei C, Kumar A, et al (2022b) Beyond separability: analyzing the linear transferability of contrastive representations to related subpopulations. arXiv preprint https://doi.org/10.48550/arXiv.2204.02683
Hayat MA, Stein G, Harrington P et al (2021) Self-supervised representation learning for astronomical images. Astrophys J Lett 911(2):L33. https://doi.org/10.3847/2041-8213/abf2c7
He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778, https://doi.org/10.1109/CVPR.2016.90
He K, Fan H, Wu Y, et al (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738, https://doi.org/10.1109/CVPR42600.2020.00975
He K, Chen X, Xie S, et al (2022) Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16000–16009, https://doi.org/10.1109/CVPR52688.2022.01553
He Y, Yang G, Ge R, et al (2023) Geometric visual similarity learning in 3d medical image self-supervised pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9538–9547, https://doi.org/10.1109/CVPR52729.2023.00920
Heim E (2023) Towards better understanding of domain shift on linear-probed visual foundation models. In: NeurIPS 2023 Workshop ICBINB
Helber P, Bischke B, Dengel A et al (2019) Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE J Select Top Appl Earth Observ Remote Sens 12(7):2217–2226. https://doi.org/10.1109/JSTARS.2019.2918242
Hendrycks D, Mazeika M, Kadavath S, et al (2019) Using self-supervised learning can improve model robustness and uncertainty. Adv Neural Inform Process Syst 32
Hermann K, Lampinen A (2020) What shapes feature representations? Exploring datasets, architectures, and training. arXiv preprint https://doi.org/10.48550/arXiv.2006.12433
Hinojosa C, Liu S, Ghanem B (2024) Colormae: Exploring data-independent masking strategies in masked autoencoders. In: European conference on computer vision, Springer, pp 432–449, https://doi.org/10.1007/978-3-031-72661-3_25
Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. arXiv preprint https://doi.org/10.48550/arXiv.2006.11239
Hondru V, Croitoru FA, Minaee S et al (2025) Masked image modeling: a survey. Int J Comput Vis. https://doi.org/10.1007/s11263-025-02524-1
Hong GZ, Cui Y, Fuxman A, et al (2023) Towards understanding the effect of pretraining label granularity. arXiv preprint https://doi.org/10.48550/arXiv.2303.16887
Hou L, Shen H, Cao Q et al (2021) Self-supervised GANs with label augmentation. Adv Neural Inf Process Syst 34:13019–13031
Hou Z, Sun F, Chen YK, et al (2022) Milan: Masked image pretraining on language assisted representation. arXiv preprint https://doi.org/10.48550/arXiv.2208.06049
Hu H, Wang X, Zhang Y et al (2024) A comprehensive survey on contrastive learning. Neurocomputing 610:128645. https://doi.org/10.1016/j.neucom.2024.128645
Hu B, Wang X, Liu W (2025) Personvit: large-scale self-supervised vision transformer for person re-identification. Mach Vis Appl 36(2):32. https://doi.org/10.1007/s00138-025-01659-y
Huang W, Yi M, Zhao X (2021) Towards the generalization of contrastive self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2111.00743
Huang G, Laradji I, Vazquez D et al (2022) A survey of self-supervised and few-shot object detection. IEEE Trans Pattern Anal Mach Intell 45(4):4071–4089. https://doi.org/10.1109/TPAMI.2022.3199617
Huang SC, Pareek A, Jensen M et al (2023) Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digital Med 6(1):74. https://doi.org/10.1038/s41746-023-00811-0
Huang Z, Jiang R, Aeron S, et al (2024a) Systematic comparison of semi-supervised and self-supervised learning for medical image classification. In: 2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR), IEEE, pp 22282–22293, https://doi.org/10.1109/CVPR52733.2024.02103
Huang Z, Jin X, Lu C et al (2024b) Contrastive masked autoencoders are stronger vision learners. IEEE Trans Pattern Anal Mach Intell 46(4):2506–2517. https://doi.org/10.1109/TPAMI.2023.3336525

Hua T, Wang W, Xue Z, et al (2021) On feature decorrelation in self-supervised learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9598–9608, https://doi.org/10.1109/ICCV48922.2021.00946

Hua T, Tian Y, Ren S, et al (2022) Self-supervision through random segments with autoregressive coding (randsac). arXiv preprint https://doi.org/10.48550/arXiv.2203.12054

Hu Z, Dong Y, Wang K, et al (2020) Gpt-gnn: Generative pre-training of graph neural networks. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp 1857–1867, https://doi.org/10.1145/3394486.3403237

Hu Q, Wang X, Hu W, et al (2021) 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, https://doi.org/10.1109/CVPR46437.2021.00113

Hu R, Debnath S, Xie S, et al (2022) Exploring long-sequence masked autoencoders. arXiv preprint https://doi.org/10.48550/arXiv.2210.07224

Hu F, Zhang C, Guo J, et al (2024) An asymmetric augmented self-supervised learning method for unsupervised fine-grained image hashing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 17648–17657, https://doi.org/10.1109/CVPR52733.2024.01671

Huynh T, Kornblith S, Walter MR, et al (2022) Boosting contrastive self-supervised learning with false negative cancellation. In: 2022 IEEE/CVF winter conference on applications of computer vision (WACV), IEEE, pp 986–996, https://doi.org/10.1109/WACV51458.2022.00106

Ibrahim M, Bouchacourt D, Morcos A (2022) Robust self-supervised learning with lie groups. arXiv preprint https://doi.org/10.48550/arXiv.2210.13356

Imran AAZ, Hatamizadeh A, Ananth SP, et al (2018) Automatic segmentation of pulmonary lobes using a progressive dense v-network. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer International Publishing, pp 282–290, https://doi.org/10.1007/978-3-030-00889-5_32

Irvin J, Rajpurkar P, Ko M, et al (2019) Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence, pp 590–597, https://doi.org/10.1609/aaai.v33i01.3301590

Jaeger S, Candemir S, Antani S et al (2014) Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg 4(6):475. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20

Jaiswal A, Babu AR, Zadeh MZ et al (2020) A survey on contrastive self-supervised learning. Technologies 9(1):2. https://doi.org/10.3390/technologies9010002

Jang J, Kim S, Yoo K, et al (2023) Self-distilled self-supervised representation learning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2829–2839, https://doi.org/10.1109/WACV56688.2023.00285

Jarca A, Croitoru FA, Ionescu RT (2024) Cbm: Curriculum by masking. arXiv:2407.05193https://doi.org/10.48550/arXiv.2407.05193

Ji X, Henriques JF, Vedaldi A (2019) Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9865–9874, https://doi.org/10.1109/ICCV.2019.00996

Ji H, Gao Z, Zhang Y et al (2022) Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks. IEEE Trans Geosci Remote Sens 60:1–13. https://doi.org/10.1109/TGRS.2022.3184080

Ji S, Han S, Rhee J (2023) Multi-view masked autoencoder for general image representation. Appl Sci 13(22):12413. https://doi.org/10.3390/app132212413

Jia J, Chan PK (2022) Feature decoupling in self-supervised representation learning for open set recognition. arXiv preprint https://doi.org/10.48550/arXiv.2209.14385

Jia C, Yang Y, Xia Y, et al (2021) Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint https://doi.org/10.48550/arXiv.2102.05918

Jiang J, Veeraraghavan H (2024) Self-supervised pretraining in the wild imparts image acquisition robustness to medical image transformers: an application to lung cancer segmentation. Proc Mach Learn Res 250:708

Jiang J, Shu Y, Wang J, et al (2022a) Transferability in deep learning: a survey. arXiv preprint https://doi.org/10.48550/arXiv.2201.05867

Jiang J, Tyagi N, Tringale K, et al (2022b) Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (SMIT). In: International conference on medical image computing and computer-assisted intervention, Springer, pp 556–566, https://doi.org/10.1007/978-3-031-16440-8_53

Jiang Z, Chen Y, Liu M, et al (2023) Layer grafted pre-training: bridging contrastive learning and masked image modeling for label-efficient representations. arXiv preprint https://doi.org/10.48550/arXiv.2302.14138

Jiang J, Rangnekar A, Veeraraghavan H (2025) Co-distilled attention guided masked image modeling with noisy teacher for self-supervised learning on medical images. In: Medical imaging with deep learning
Jin W, Derr T, Liu H, et al (2020) Self-supervised learning on graphs: deep insights and new direction. arXiv preprint https://doi.org/10.48550/arXiv.2006.10141
Jing L, Tian Y (2020) Self-supervised visual feature learning with deep neural networks: a survey. IEEE Trans Pattern Anal Mach Intell 43(11):4037–4058. https://doi.org/10.1109/TPAMI.2020.2992393
Jing L, Yang X, Liu J, et al (2018) Self-supervised spatiotemporal feature learning via video rotation prediction. arXiv preprint https://doi.org/10.48550/arXiv.1811.11387
Jing L, Vincent P, LeCun Y, et al (2021) Understanding dimensional collapse in contrastive self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2110.09348
Johnson J, Hariharan B, Van Der Maaten L, et al (2017) Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2901–2910, https://doi.org/10.1109/CVPR.2017.215
Jonske F, Kim M, Nasca E et al (2025) Why does my medical ai look at pictures of birds? Exploring the efficacy of transfer learning across domain boundaries. Comput Methods Programs Biomed 261:108634. https://doi.org/10.1016/j.cmpb.2025.108634
Kakogeorgiou I, Gidaris S, Psomas B, et al (2022) What to hide from your students: attention-guided masked image modeling. In: European conference on computer vision, Springer, pp 300–318, https://doi.org/10.1007/978-3-031-20056-4_18
Kang M, Song H, Park S, et al (2023) Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3344–3354, https://doi.org/10.1109/CVPR52729.2023.00326
Kaplan J, McCandlish S, Henighan T, et al (2020) Scaling laws for neural language models. arXiv preprint https://doi.org/10.48550/arXiv.2001.08361
Kather JN, Weis CA, Bianconi F et al (2016) Multi-class texture analysis in colorectal cancer histology. Sci Rep 6(1):1–11. https://doi.org/10.1038/srep27988
Kather JN, Halama N, Marx A (2018) 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo. https://doi.org/10.5281/zenodo.1214456
Kawashti YA, Khattab D, Aref MM (2023) Exploring self-supervised pretraining datasets for complex scene understanding. Int J Intell Comput Inform Sci 23(2):62–73. https://doi.org/10.21608/ijicis.2023.193520.1255
Kendall A, Gal Y, Cipolla R (2018) Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7482–7491, https://doi.org/10.1109/CVPR.2018.00781
Kermany DS, Goldbaum M, Cai W et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172(5):1122–1131. https://doi.org/10.1016/j.cell.2018.02.010
Khan MO, Fang Y (2023) Revisiting fine-tuning strategies for self-supervised medical imaging analysis. arXiv preprint https://doi.org/10.48550/arXiv.2307.10915
Khan A, AlBarri S, Manzoor MA (2022) Contrastive self-supervised learning: a survey on different architectures. In: 2022 2nd International conference on artificial intelligence (icai), IEEE, pp 1–6, https://doi.org/10.1109/ICAI55435.2022.9773725
Khan A, Asmatullah L, Malik A, et al (2025a) A survey on self-supervised contrastive learning for multimodal text-image analysis. arXiv preprint https://doi.org/10.48550/arXiv.2503.11101
Khan A, Sohail A, Fiaz M, et al (2025b) A survey of the self supervised learning mechanisms for vision transformers. arXiv preprint https://doi.org/10.48550/arXiv.2408.17059
Kim D, Cho D, Yoo D, et al (2018) Learning image representations by completing damaged jigsaw puzzles. In: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 793–802, https://doi.org/10.1109/WACV.2018.00092
Kirillov A, Mintun E, Ravi N, et al (2023) Segment anything. In: 2023 IEEE/CVF international conference on computer vision (ICCV), IEEE, pp 3992–4003, https://doi.org/10.1109/CVPR52733.2024.00300
Kolesnikov A, Zhai X, Beyer L (2019) Revisiting self-supervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1920–1929, https://doi.org/10.1109/CVPR.2019.00202
Kolesnikov A, Beyer L, Zhai X, et al (2020) Big transfer (bit): General visual representation learning. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, pp 491–507, https://doi.org/10.1007/978-3-030-58558-7_29
Kong X, Zhang X (2023) Understanding masked image modeling via learning occlusion invariant feature. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6241–6251, https://doi.org/10.1109/CVPR52729.2023.00604

Kong L, Ma MQ, Chen G, et al (2023) Understanding masked autoencoders via hierarchical latent variable models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7918–7928, https://doi.org/10.1109/CVPR52729.2023.00765
Konstantakos S, Cani J, Mademlis I et al (2025) Self-supervised visual learning in the low-data regime: a comparative evaluation. Neurocomputing 620:129199. https://doi.org/10.1016/j.neucom.2024.129199
Kornblith S, Chen T, Lee H, et al (2021) Why do better loss functions lead to less transferable features? arXiv preprint https://doi.org/10.48550/arXiv.2010.16402
Krause J, Stark M, Deng J, et al (2013) 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops, pp 554–561, https://doi.org/10.1109/ICCVW.2013.77
Krishna K, Garg S, Bigham JP, et al (2023) Downstream datasets make surprisingly good pretraining corpora. arXiv preprint https://doi.org/10.48550/arXiv.2209.14389
Krishnan R, Rajpurkar P, Topol EJ (2022) Self-supervised learning in medicine and healthcare. Nat Biomed Eng. https://doi.org/10.1038/s41551-022-00914-1
Krizhevsky A (2009) Learning multiple layers of features from tiny images. University of Toronto, ON, Canada
Kumar A, Raghunathan A, Jones RM, et al (2022a) Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint https://doi.org/10.48550/arXiv.2202.10054
Kumar P, Rawat P, Chauhan S (2022b) Contrastive self-supervised learning: review, progress, challenges and future research directions. Int J Multimedia Inform Retrieval. https://doi.org/10.1007/s13735-022-00245-6
Kuznetsova A, Rom H, Alldrin N et al (2020) The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. Int J Comput Vis 128(7):1956–1981. https://doi.org/10.1007/s11263-020-01316-z
Kwasigroch A, Grochowski M, Mikołajczyk A (2020) Self-supervised learning to increase the performance of skin lesion classification. Electronics 9(11):1930. https://doi.org/10.3390/electronics9111930
Lahrichi S, Sheng Z, Xia S, et al (2025) Is self-supervised pre-training on satellite imagery better than imagenet? A systematic study with sentinel-2. arXiv preprint https://doi.org/10.48550/arXiv.2502.10669
Larsson G, Maire M, Shakhnarovich G (2016) Learning representations for automatic colorization. In: Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, pp 577–593, https://doi.org/10.1007/978-3-319-46493-0_35
Lavoie MA, Mahmoud A, Waslander SL (2025) Large self-supervised models bridge the gap in domain adaptive object detection. In: Proceedings of the computer vision and pattern recognition conference, pp 4692–4702
Ledig C, Theis L, Huszár F, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690, https://doi.org/10.1109/CVPR.2017.19
Lee H, Hwang SJ, Shin J (2020) Self-supervised label augmentation via input transformations. https://doi.org/10.48550/arXiv.1910.05872
Lee JD, Lei Q, Saunshi N et al (2021) Predicting what you already know helps: provable self-supervised learning. Adv Neural Inf Process Syst 34:309–323. https://doi.org/10.3390/rs12081246
Lee N, Lee J, Park C (2022a) Augmentation-free self-supervised learning on graphs. In: Proceedings of the AAAI conference on artificial intelligence, pp 7372–7380, https://doi.org/10.1609/aaai.v36i7.20700
Lee Y, Willette J, Kim J, et al (2022b) Exploring the role of mean teachers in self-supervised masked autoencoders. arXiv preprint https://doi.org/10.48550/arXiv.2210.02077
Lee Y, Chen AS, Tajwar F, et al (2023a) Surgical fine-tuning improves adaptation to distribution shifts. arXiv preprint https://doi.org/10.48550/arXiv.2210.11466
Lee JH, Yoon D, Ji B, et al (2023b) Rethinking evaluation protocols of visual representations learned via self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2304.03456
Lee M, Park S, Heo B, et al (2024) Seit++: Masked token modeling improves storage-efficient training. In: European conference on computer vision, Springer, pp 180–197, https://doi.org/10.1007/978-3-031-73390-1_11
Lehner J, Alkin B, Fürst A, et al (2024) Contrastive tuning: a little help to make masked autoencoders forget. In: Proceedings of the AAAI conference on artificial intelligence, pp 2965–2973, https://doi.org/10.1609/aaai.v38i4.28078
Leminen Madsen S, Mathiassen SK, Dyrmann M et al (2020) Open plant phenotype database of common weeds in Denmark. Remote Sens 12(8):1246. https://doi.org/10.3390/rs12081246
Le Y, Yang X (2015) Tiny imagenet visual recognition challenge. CS 231N 7(7):3. https://doi.org/10.57702/19nlfifl

Li LJ, Fei-Fei L (2007) What, where and who? Classifying events by scene and object recognition. In: 2007 IEEE 11th international conference on computer vision, IEEE, pp 1–8, https://doi.org/10.1109/ICCV.2007.4408872
Li S, Tang H (2025) Multimodal alignment and fusion: a survey. arXiv preprint https://doi.org/10.48550/arXiv.2411.17040
Li X, Xiong H, An H, et al (2020) Rifle: Backpropagation in depth for deep transfer learning through re-initializing the fully-connected layer. In: International conference on machine learning, PMLR, pp 6010–6019
Li Z, Chen Z, Yang F et al (2021a) Mst: Masked self-supervised transformer for visual representation. Adv Neural Inf Process Syst 34:13165–13176
Li H, Park S, Dara A, et al (2021b) Neural model robustness for skill routing in large-scale conversational ai systems: a design choice exploration. arXiv preprint https://doi.org/10.48550/arXiv.2103.03373
Li H, Tao R, Li J, et al (2021c) Multi-pretext attention network for few-shot learning with self-supervision. In: 2021 IEEE international conference on multimedia and expo (ICME), IEEE, pp 1–6, https://doi.org/10.1109/ICME51207.2021.9428447
Li C, Yang J, Zhang P, et al (2021d) Efficient self-supervised vision transformers for representation learning. arXiv preprint https://doi.org/10.48550/arXiv.2106.09785
Li J, Zhou P, Xiong C, et al (2021e) Prototypical contrastive learning of unsupervised representations. arXiv preprint https://doi.org/10.48550/arXiv.2005.04966
Li FF, Andreeto M, Ranzato M, et al (2022a) Caltech 101. https://doi.org/10.22002/D1.20086
Li X, Ge Y, Yi K, et al (2022b) mc-beit: Multi-choice discretization for image bert pre-training. In: European conference on computer vision, Springer, pp 231–246, https://doi.org/10.1007/978-3-031-20056-4_14
Li G, Zheng H, Liu D et al (2022c) Semmae: Semantic-guided masking for learning masked autoencoders. Adv Neural Inf Process Syst 35:14290–14302
Li X, Wang W, Yang L, et al (2022d) Uniform masking: enabling MAE pre-training for pyramid-based vision transformers with locality. arXiv preprint https://doi.org/10.48550/arXiv.2205.10063
Li J, Wang Y, Zhang X, et al (2022e) Progressively compressed auto-encoder for self-supervised representation learning. In: The eleventh international conference on learning representations
Li Z, Zhu Y, Yang F, et al (2022f) 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, https://doi.org/10.1109/CVPR52688.2022.01422
Li T, Chang H, Mishra S, et al (2023a) Mage: Masked generative encoder to unify representation learning and image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2142–2152, https://doi.org/10.1109/CVPR52729.2023.00213
Li Y, Fan H, Hu R, et al (2023b) Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 23390–23400, https://doi.org/10.1109/CVPR52729.2023.02240
Li J, Li D, Savarese S, et al (2023c) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning, PMLR, pp 19730–19742
Li S, Wu D, Wu F, et al (2023d) Architecture-agnostic masked image modeling–from ViT back to CNN. arXiv preprint https://doi.org/10.48550/arXiv.2205.13943
Li S, Zhang L, Wang Z, et al (2024a) Masked modeling for self-supervised representation learning on vision and beyond. arXiv preprint https://doi.org/10.48550/arXiv.2401.00897
Li Z, Zhu Y, Chen Z, et al (2024b) Efficient masked autoencoders with self-consistency. arXiv preprint https://doi.org/10.48550/arXiv.2302.14431
Li X, Huang J, Sun G, et al (2025) Self-supervised learning for mri reconstruction: a review and new perspective. Magn Reson Mater Phys, Biol Med pp 1–22. https://doi.org/10.1007/s10334-025-01274-y
Liang PP, Deng Z, Ma M, et al (2023) Factorized contrastive learning: going beyond multi-view redundancy. arXiv preprint https://doi.org/10.48550/arXiv.2306.05268
Liang PP, Ling CK, Cheng Y, et al (2024) Multimodal learning without labeled multimodal data: guarantees and applications. arXiv preprint https://doi.org/10.48550/arXiv.2306.04539
Lim S, Kim I, Kim T, et al (2019) Fast autoaugment. Adv Neural Inform Process Syst 32
Lin TY, Maire M, Belongie S, et al (2014) Microsoft COCO: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, pp 740–755, https://doi.org/10.1007/978-3-319-10602-1_48
Lin T, Zhang J, Xu Y, et al (2024) Relational contrastive learning and masked image modeling for scene text recognition. arXiv preprint https://doi.org/10.48550/arXiv.2411.11219
Liu Y, Jain A, Eng C et al (2020) A deep learning system for differential diagnosis of skin diseases. Nat Med 26(6):900–908. https://doi.org/10.1038/s41591-020-0842-3

Liu H, HaoChen JZ, Gaidon A, et al (2021a) Self-supervised learning is more robust to dataset imbalance. arXiv preprint https://doi.org/10.48550/arXiv.2110.05025
Liu S, Li Z, Sun J (2021b) Self-emd: Self-supervised object detection without imagenet. arXiv preprint https://doi.org/10.48550/arXiv.2011.13677
Liu KJ, Suganuma M, Okatani T (2022a) Bridging the gap from asymmetry tricks to decorrelation principles in non-contrastive self-supervised learning. Adv Neural Inf Process Syst 35:19824–19835
Liu Y, Jin M, Pan S et al (2022b) Graph self-supervised learning: a survey. IEEE Trans Knowl Data Eng 35(6):5879–5900. https://doi.org/10.1109/TKDE.2022.3172903
Liu Z, Li B, Han C, et al (2022c) Masked reconstruction contrastive learning with information bottleneck principle. arXiv preprint https://doi.org/10.48550/arXiv.2211.09013
Liu J, Huang X, Zheng J, et al (2023a) Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6252–6261, https://doi.org/10.1109/CVPR52729.2023.00605
Liu H, Jiang X, Li X, et al (2023b) The devil is in the frequency: geminated gestalt autoencoder for self-supervised visual pre-training. In: Proceedings of the AAAI conference on artificial intelligence, pp 1649–1656, https://doi.org/10.1609/aaai.v37i2.25252
Liu Y, Zhang S, Chen J, et al (2023c) Improving pixel-based MIM by reducing wasted modeling capability. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5361–5372, https://doi.org/10.1109/ICCV51070.2023.00494
Liu Y, Zhang S, Chen J, et al (2023d) Pixmim: Rethinking pixel reconstruction in masked image modeling. arXiv preprint https://doi.org/10.48550/arXiv.2303.02416
Liu X, Zhang F, Hou Z et al (2023e) Self-supervised learning: generative or contrastive. IEEE Trans Knowl Data Eng 35(1):857–876. https://doi.org/10.1109/TKDE.2021.3090866
Liu X, Zhou J, Kong T, et al (2023f) Exploring target representations for masked autoencoders. arXiv preprint https://doi.org/10.48550/arXiv.2209.03917
Liu X, Miao X, Jiang H et al (2024a) Tower masking mim: a self-supervised pretraining method for power line inspection. IEEE Trans Ind Inf 20(1):513–523. https://doi.org/10.1109/TII.2023.3268479
Liu Z, Alavi A, Li M, et al (2024b) Self-supervised learning for time series: contrastive or generative? arXiv preprint https://doi.org/10.48550/arXiv.2403.09809
Liu Z, Kainth K, Zhou A et al (2024c) A review of self-supervised, generative, and few-shot deep learning methods for data-limited magnetic resonance imaging segmentation. NMR Biomed 37(8):e5143. https://doi.org/10.1002/nbm.5143
Lučić M, Tschannen M, Ritter M, et al (2019) High-fidelity image generation with fewer labels. In: International conference on machine learning, PMLR, pp 4183–4192
Lu J, Clark C, Zellers R, et al (2022) Unified-io: A unified model for vision, language, and multi-modal tasks. In: The eleventh international conference on learning representations
Lu CZ, Jin X, Hou Q, et al (2023) Delving deeper into data scaling in masked image modeling. arXiv preprint https://doi.org/10.48550/arXiv.2305.15248
Luo Y, Ren M, Zhang SQ (2023) BIM: Block-wise self-supervised learning with masked image modeling. arXiv preprint https://doi.org/10.48550/arXiv.2311.17218
Ma S, Jw L (2023) Self-supervised contrastive learning for heterogeneous graph based on multi-pretext tasks. Neural Comput Appl 35(14):10275–10296. https://doi.org/10.1007/s00521-023-08234-4
Ma J, Hu T, Wang W (2023) Deciphering the projection head: representation evaluation self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2301.12189
Ma X, Liu C, Xie C et al (2024) Disjoint masking with joint distillation for efficient masked image modeling. IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2023.3306840
Maitra A, Mitra S, Manna S et al (2025) Decorrelation-based self-supervised visual representation learning for writer identification. ACM Trans Asian Low-Resour Lang Inform Process 24(7):1–17. https://doi.org/10.1145/3746062
Maji S, Rahtu E, Kannala J, et al (2013) Fine-grained visual classification of aircraft. arXiv preprint https://doi.org/10.48550/arXiv.1306.5151
Maninis KK, Chen K, Ghosh S, et al (2025) Tips: Text-image pretraining with spatial awareness. arXiv preprint https://doi.org/10.48550/arXiv.2410.16512
Manna S, Chattopadhyay S, Dey R, et al (2024) Dystress: Dynamically scaled temperature in self-supervised contrastive learning. arXiv preprint https://doi.org/10.48550/arXiv.2308.01140
Mao HH (2020) A survey on self-supervised pre-training for sequential transfer learning in neural networks. arXiv preprint https://doi.org/10.48550/arXiv.2007.00800
Mao J, Yin X, Chang Y, et al (2022) Improvements to self-supervised representation learning for masked image modeling. arXiv preprint https://doi.org/10.48550/arXiv.2205.10546
Mao J, Zhou H, Yin X, et al (2023) Masked autoencoders are effective solution to transformer data-hungry. arXiv preprint https://doi.org/10.48550/arXiv.2212.05677

Marchetti GL, Tegnér G, Varava A, et al (2023) Equivariant representation learning via class-pose decomposition. In: International conference on artificial intelligence and statistics, PMLR, pp 4745–4756
Marks M, Knott M, Kondapaneni N et al (2025) A closer look at benchmarking self-supervised pre-training with image classification. Int J Comput Vis. https://doi.org/10.1007/s11263-025-02402-w
Markus AF, Kors JA, Rijnbeek PR (2021) The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 113:103655. https://doi.org/10.1016/j.jbi.2020.103655
Martinović I, Mao S, Dousty M et al (2025) X-ray modalities in the era of artificial intelligence: overview of self-supervised learning approach. Facets 10:1–17. https://doi.org/10.1139/facets-2024-0229
Matsoukas C, Haslum JF, Sorkhei M, et al (2022) What makes transfer learning work for medical images: feature reuse & other factors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9225–9234, https://doi.org/10.48550/arXiv.2203.01825
Medina C, Devos A, Grossglauser M (2020) Self-supervised prototypical transfer learning for few-shot classification. arXiv preprint https://doi.org/10.48550/arXiv.2006.11325
Meehan C, Bordes F, Vincent P et al (2023) Do SSL models have déjà vu? A case of unintended memorization in self-supervised learning. Adv Neural Inf Process Syst 36:42775–42798
Mensink T, Uijlings J, Kuznetsova A et al (2022) Factors of influence for transfer learning across diverse appearance domains and task types. IEEE Trans Pattern Anal Mach Intell 44(12):9298–9314. https://doi.org/10.1109/TPAMI.2021.3129870
Mialon G, Balestriero R, LeCun Y (2024) Variance covariance regularization enforces pairwise independence in self-supervised representations. arXiv preprint https://doi.org/10.48550/arXiv.2209.14905
Minderer M, Bachem O, Houlsby N, et al (2020) Automatic shortcut removal for self-supervised representation learning. arXiv preprint https://doi.org/10.48550/arXiv.2002.08822
Mishra S, Robinson J, Chang H, et al (2022) A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. arXiv preprint https://doi.org/10.48550/arXiv.2210.16870
Misra I, Maaten Lvd (2020) Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6707–6717, https://doi.org/10.1109/CVPR42600.2020.00674
Mitrovic J, McWilliams B, Walker J, et al (2020) Representation learning via invariant causal mechanisms. arXiv preprint arXiv:2010.07922
Mi L, Xu C, Castillo-Navarro J, et al (2024) Congeo: Robust cross-view geo-localization across ground view variations. In: European conference on computer vision, Springer, pp 214–230, https://doi.org/10.1007/978-3-031-72630-9_13
Mizrahi D, Bachmann R, Kar OF, et al (2023) 4m: Massively multimodal masked modeling. arXiv preprint https://doi.org/10.48550/arXiv.2312.06647
Moher D, Liberati A, Tetzlaff J et al (2010) Preferred reporting items for systematic reviews and meta-analyses: the Prisma statement. Int J Surg 8(5):336–341
Mojvzivsek D, Huula J, Li Z, et al (2025) Neural approaches to sat solving: design choices and interpretability. arXiv preprint https://doi.org/10.48550/arXiv.2504.01173
Moon W, Kim JH, Heo JP (2022) Tailoring self-supervision for supervised learning. In: Computer vision–ECCV 2022: 17th European conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV, Springer, pp 346–364, https://doi.org/10.1007/978-3-031-19806-9_20
Morningstar W, Bijamov A, Duvarney C, et al (2024) Augmentations vs algorithms: what works in self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2403.05726
Mo S, Sun Z, Li C (2023) Multi-level contrastive learning for self-supervised vision transformers. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2778–2787, https://doi.org/10.1109/WACV56688.2023.00280
Moutakanni T, Oquab M, Szafraniec M, et al (2024) You don't need data-augmentation in self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2406.09294
Mundhenk TN, Ho D, Chen BY (2018) Improvements to context based self-supervised learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9339–9348, https://doi.org/10.1109/CVPR.2018.00973
Navarro F, Watanabe C, Shit S, et al (2021) Evaluating the robustness of self-supervised learning in medical imaging. arXiv preprint https://doi.org/10.48550/arXiv.2105.06986
Nedungadi V, Kariryaa A, Oehmcke S, et al (2024) Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. arXiv preprint https://doi.org/10.48550/arXiv.2405.02771
Newell A, Deng J (2020) How useful is self-supervised pretraining for visual tasks? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7345–7354, https://doi.org/10.1109/CVPR42600.2020.00737
Newton A, Buckeye A, Nih AA, et al (2015) Second annual data science bowl. https://kaggle.com/competitions/second-annual-data-science-bowl

Neyshabur B, Sedghi H, Zhang C (2021) What is being transferred in transfer learning? arXiv preprint https://doi.org/10.48550/arXiv.2008.11687
Nilsback ME, Zisserman A (2008) Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing, IEEE, pp 722–729, https://doi.org/10.1109/ICVGIP.2008.47
Noroozi M, Favaro P (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision, Springer, pp 69–84, https://doi.org/10.1007/978-3-319-46466-4_5
Noroozi M, Vinjimoor A, Favaro P, et al (2018) Boosting self-supervised learning via knowledge transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9359–9367, https://doi.org/10.1109/CVPR.2018.00975
Ogidi FC, Eramian MG, Stavness I (2023) Benchmarking self-supervised contrastive learning methods for image-based plant phenotyping. Plant Phenomics 5:0037. https://doi.org/10.34133/plantphenomics.0037
Ohri K, Kumar M (2021) Review on self-supervised image recognition using deep neural networks. Knowl-Based Syst 224:107090. https://doi.org/10.1016/j.knosys.2021.107090
Ong CA, Tay FS, Then YL et al (2025) An evaluation of a pre-trained transformer-based self-distillation model (dinov2) for cross-domain plant species identification. Neural Comput Appl 37(26):21969–21995
Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748
Oquab M, Darcet T, Moutakanni T, et al (2024) Dinov2: Learning robust visual features without supervision. arXiv preprint https://doi.org/10.48550/arXiv.2304.07193
Ouyang C, Biffi C, Chen C, et al (2020) Self-supervision with superpixels: training few-shot medical image segmentation without annotation. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, Proceedings, Part XXIX 16, Springer, pp 762–780, https://doi.org/10.1007/978-3-030-58526-6_45
Ozbulak U, Lee HJ, Boga B, et al (2023) Know your self-supervised learning: a survey on image-based generative and discriminative training. arXiv preprint https://doi.org/10.48550/arXiv.2305.13689
Pal A, Balasubramanian VN (2018) Adversarial data programming: using gans to relax the bottleneck of curated labeled data. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1556–1565, https://doi.org/10.1109/CVPR.2018.00168
Pal DK, Nallamothu S, Savvides M (2019) Towards a hypothesis on visual transformation based self-supervision. arXiv preprint https://doi.org/10.48550/arXiv.1911.10594
Pan H, Guo Y, Deng Q et al (2023a) Improving fine-tuning of self-supervised models with contrastive initialization. Neural Netw 159:198–207. https://doi.org/10.1016/j.neunet.2022.12.012
Pan H, Liu C, Wang W, et al (2023b) Img2vec: A teacher of high token-diversity helps masked autoencoders. arXiv preprint https://doi.org/10.48550/arXiv.2304.12535
Pani K, Chawla I (2025) Examining the quality of learned representations in self-supervised medical image analysis: a comprehensive review and empirical study. Multimedia Tools Appl 84(9):6295–6325. https://doi.org/10.1007/s11042-024-19072-4
Park W, Ryu J (2024) Fine-grained self-supervised learning with jigsaw puzzles for medical image classification. Comput Biol Med 174:108460. https://doi.org/10.1016/j.compbiomed.2024.108460
Park T, Efros AA, Zhang R, et al (2020) Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision, Springer, pp 319–345, https://doi.org/10.1007/978-3-030-58545-7_19
Park JY, Biza O, Zhao L, et al (2022) Learning symmetric embeddings for equivariant world models. arXiv preprint https://doi.org/10.48550/arXiv.2204.11371
Park N, Kim W, Heo B, et al (2023) What do self-supervised vision transformers learn? arXiv preprint https://doi.org/10.48550/arXiv.2305.00729
Parkhi OM, Vedaldi A, Zisserman A, et al (2012) Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp 3498–3505, https://doi.org/10.1109/CVPR.2012.6248092
Pathak D, Krahenbuhl P, Donahue J, et al (2016) Context encoders: feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2536–2544, https://doi.org/10.1109/CVPR.2016.278
Peng Z, Dong L, Bao H, et al (2022) Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint https://doi.org/10.48550/arXiv.2208.06366
Pototzky D, Sultan A, Schmidt-Thieme L (2022) Fastsiam: Resource-efficient self-supervised learning on a single gpu. In: DAGM German conference on pattern recognition, Springer, pp 53–67
Poursaeed O, Jiang T, Qiao H, et al (2020) Self-supervised learning of point clouds via orientation estimation. In: 2020 International conference on 3D vision (3DV), IEEE, pp 1018–1028, https://doi.org/10.1109/3DV50981.2020.00112
Przewikeźlikowski M, Pyla M, Zieliński B et al (2024) Augmentation-aware self-supervised learning with conditioned projector. Knowl-Based Syst 305:112572. https://doi.org/10.1016/j.knosys.2024.112572

Purushwalkam S, Gupta A (2020) Demystifying contrastive self-supervised learning: invariances, augmentations and dataset biases. arXiv preview https://doi.org/10.48550/arXiv.2007.13916
Qian Q, Xu Y, Hu J, et al (2022) 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
Qu L, Liu S, Liu X et al (2022) Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis. Phys Med Biol 67(20):20TR01. https://doi.org/10.1088/1361-6560/ac910a
Quan S, Hirano M, Yamakawa Y (2023) Semantic information in contrastive learning. In: 2023 IEEE/CVF international conference on computer vision (ICCV), IEEE, pp 5663–5673, https://doi.org/10.1109/ICCV51070.2023.00523
Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR, pp 8748–8763
Raghu M, Zhang C, Kleinberg J, et al (2019) Transfusion: Understanding transfer learning for medical imaging. arXiv preprint https://doi.org/10.48550/arXiv.1902.07208
Ramesh A, Pavlov M, Goh G, et al (2021) Zero-shot text-to-image generation. In: International conference on machine learning, PMLR, pp 8821–8831
Rani V, Nabi ST, Kumar M et al (2023) Self-supervised learning: a succinct review. Arch Comput Methods Eng 30(4):2761–2775. https://doi.org/10.1007/s11831-023-09884-2
Rani V, Kumar M, Gupta A et al (2024) Self-supervised learning for medical image analysis: a comprehensive review. Evol Syst 15(4):1607–1633. https://doi.org/10.1007/s12530-024-09581-w
Reed CJ, Yue X, Nrusimha A, et al (2022) Self-supervised pretraining improves self-supervised pretraining. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2584–2594, https://doi.org/10.1109/WACV51458.2022.00112
Ren S, Wei F, Zhang Z, et al (2023) Tinymim: An empirical study of distilling mim pre-trained models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3687–3697, https://doi.org/10.1109/CVPR52729.2023.00359
Ren S, Wang Z, Zhu H, et al (2024) Rejuvenating image-GPT as strong visual representation learners. arXiv preprint https://doi.org/10.48550/arXiv.2312.02147
Ren S, Wei F, Zhang SAZ, et al (2025) Deepmim: Deep supervision for masked image modeling. In: 2025 IEEE/CVF winter conference on applications of computer vision (WACV), IEEE, pp 879–888, https://doi.org/10.1109/WACV61041.2025.00095
Richemond PH, Tam A, Tang Y, et al (2023) The edge of orthogonality: a simple view of what makes byol tick. In: International conference on machine learning, PMLR, pp 29063–29081
Ridnik T, Ben-Baruch E, Noy A, et al (2021) Imagenet-21k pretraining for the masses. arXiv preprint https://doi.org/10.48550/arXiv.2104.10972
Risojević V, Stojnić V (2022) Do we still need imagenet pre-training in remote sensing scene classification? arXiv preprint https://doi.org/10.48550/arXiv.2111.03690
Robinson J, Chuang CY, Sra S, et al (2021a) Contrastive learning with hard negative samples. arXiv preprint https://doi.org/10.48550/arXiv.2010.04592
Robinson J, Sun L, Yu K et al (2021b) Can contrastive learning avoid shortcut solutions? Adv Neural Inf Process Syst 34:4974–4986
Rombach R, Blattmann A, Lorenz D, et al (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10684–10695, https://doi.org/10.1109/CVPR52688.2022.01042
Russakovsky O, Deng J, Su H et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115:211–252. https://doi.org/10.1007/s11263-015-0816-y
Ryali CK, Schwab DJ, Morcos AS (2021) Characterizing and improving the robustness of self-supervised learning through background augmentations. arXiv preprint arXiv:2103.12719
Saito K, Kim D, Sclaroff S, et al (2020) Universal domain adaptation through self supervision. arXiv preprint https://doi.org/10.48550/arXiv.2002.07953
Sameni S, Jenni S, Favaro P (2023) Representation learning by detecting incorrect location embeddings. In: Proceedings of the AAAI conference on artificial intelligence, pp 9704–9713, https://doi.org/10.1609/aaai.v37i8.26160
Sammani F, Joukovsky B, Deligiannis N (2023) Visualizing and understanding contrastive learning. arXiv preprint https://doi.org/10.48550/arXiv.2206.09753
Sanchez-Fernandez AJ, Moreno-Álvarez S, Rico-Gallego JA et al (2024) Self-supervised learning on small in-domain datasets can overcome supervised learning in remote sensing. IEEE J Select Top Appl Earth Observ Remote Sens 17:12797–12810. https://doi.org/10.1109/JSTARS.2024.3421622

Sariyildiz MB, Kalantidis Y, Alahari K, et al (2023) No reason for no supervision: improved generalization in supervised models. arXiv preprint https://doi.org/10.48550/arXiv.2206.15369
Scheibenreif L, Mommert M, Borth D (2022) Contrastive self-supervised data fusion for satellite imagery. ISPRS Ann Photogram Remote Sens Spatial Inform Sci 3:705–711. https://doi.org/10.5194/isprs-annals-V-3-2022-705-2022
Scherr F, Guo Q, Moraitis T (2022) Self-supervised learning through efference copies. Adv Neural Inf Process Syst 35:4543–4557
Schiappa MC, Rawat YS, Shah M (2022) Self-supervised learning for videos: a survey. ACM Comput Surv. https://doi.org/10.1145/3577925
Schürholt K, Kostadinov D, Borth D (2022) Hyper-representations: self-supervised representation learning on neural network weights for model characteristic prediction. arXiv preprint https://doi.org/10.48550/arXiv.2110.15288
Sehwag V, Wang S, Mittal P, et al (2019) Towards compact and robust deep neural networks. arXiv preprint https://doi.org/10.48550/arXiv.1906.06110
Setio AAA, Traverso A, De Bel T et al (2017) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med Image Anal 42:1–13. https://doi.org/10.1016/j.media.2017.06.015
Sharma R, Ji K, Chen C, et al (2023) AUC-CL: A batchsize-robust framework for self-supervised contrastive representation learning. In: The twelfth international conference on learning representations
Shi Y, Siddharth N, Torr P, et al (2022) Adversarial masking for self-supervised learning. In: International conference on machine learning, PMLR, pp 20026–20040
Shi C, Sun C, Wu Y, et al (2023a) Video anomaly detection via sequentially learning multiple pretext tasks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10330–10340, https://doi.org/10.1109/ICCV51070.2023.00948
Shi B, Zhang X, Wang Y, et al (2023b) Hybrid distillation: connecting masked autoencoders with contrastive learners. arXiv preprint https://doi.org/10.48550/arXiv.2306.15876
Shurrab S, Duwairi R (2022) Self-supervised learning methods and applications in medical imaging analysis: a survey. PeerJ Comput Sci 8:e1045. https://doi.org/10.7717/peerj-cs.1045
Shwartz Ziv R, LeCun Y (2024) To compress or not to compress-self-supervised learning and information theory: a review. Entropy. https://doi.org/10.3390/e26030252
Shwartz-Ziv R, Balestriero R, LeCun Y (2022) What do we maximize in self-supervised learning? arXiv preprint arXiv:2207.10081
Silberman N, Hoiem D, Kohli P et al (2012) Indoor segmentation and support inference from rgbd images. ECCV 5(7576):746–760. https://doi.org/10.1007/978-3-642-33715-4_54
Singh M, Gustafson L, Adcock A, et al (2022) Revisiting weakly supervised pre-training of visual perception models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 804–814, https://doi.org/10.1109/CVPR52688.2022.00088
Singh M, Duval Q, Alwala KV, et al (2023) The effectiveness of MAE pre-pretraining for billion-scale pretraining. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5484–5494, https://doi.org/10.1109/ICCV51070.2023.00505
Sirotkin K, Escudero-Viñolo M, Carballeira P et al (2024) Improved transferability of self-supervised learning models through batch normalization finetuning. Appl Intell 54(22):11281–11294. https://doi.org/10.1007/s10489-024-05758-7
Smith LN, Topin N (2016) Deep convolutional neural network design patterns. arXiv preprint https://doi.org/10.48550/arXiv.1611.00847
Spencer J, Russell C, Hadfield S, et al (2022) Deconstructing self-supervised monocular reconstruction: the design decisions that matter. arXiv preprint https://doi.org/10.48550/arXiv.2208.01489
Srinath Halvagal M, Laborieux A, Zenke F (2023) Implicit variance regularization in non-contrastive ssl. Adv Neural Inf Process Syst 36:63409–63436
Su JC, Maji S, Hariharan B (2020) When does self-supervision improve few-shot learning? In: European conference on computer vision, Springer, pp 645–666, https://doi.org/10.1007/978-3-030-58571-6_38
Sumbul G, Charfuelan M, Demir B, et al (2019) Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In: IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium, IEEE, pp 5901–5904, https://doi.org/10.1109/IGARSS.2019.8900532
Sun C, Shrivastava A, Singh S, et al (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision, pp 843–852, https://doi.org/10.1109/ICCV.2017.97
Sun W, Tagliasacchi A, Deng B et al (2021) Canonical capsules: self-supervised capsules in canonical pose. Adv Neural Inf Process Syst 34:24993–25005
Taherdoost H (2024) Beyond supervised: the rise of self-supervised learning in autonomous systems. Information 15(8):491. https://doi.org/10.3390/info15080491

Taleb A, Loetzsch W, Danz N, et al (2000) 3d self-supervised methods for medical imaging. arXiv preprint https://doi.org/10.48550/arXiv.2006.03829
Tamkin A, Wu M, Goodman N (2021) Viewmaker networks: learning views for unsupervised representation learning. arXiv preprint https://doi.org/10.48550/arXiv.2010.07432
Tan Z, Yang J, Huang W, et al (2024) Information flow in self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2309.17281
Tao C, Zhu X, Su W, et al (2023) Siamese image modeling for self-supervised vision representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2132–2141, https://doi.org/10.1109/CVPR52729.2023.00212
Tayebi Arasteh S, Misera L, Kather JN et al (2024) Enhancing diagnostic deep learning via self-supervised pretraining on large-scale, unlabeled non-medical images. Eur Radiol Exp 8(1):10. https://doi.org/10.1186/s41747-023-00411-3
Tendle A, Hasan MR (2021) A study of the generalizability of self-supervised representations. Mach Learn Appl 6:100124. https://doi.org/10.1016/j.mlwa.2021.100124
Teng J, Huang W, He H (2022) Can pretext-based self-supervised learning be boosted by downstream data? A theoretical analysis. In: International conference on artificial intelligence and statistics, PMLR, pp 4198–4216
Thapa S (2022) Survey on self-supervised multimodal representation learning and foundation models. arXiv preprint https://doi.org/10.48550/arXiv.2211.15837
Thomee B, Shamma DA, Friedland G et al (2016) YFCC100M: The new data in multimedia research. Commun ACM 59(2):64–73. https://doi.org/10.1145/2812802
Tian Y, Krishnan D, Isola P (2020a) Contrastive multiview coding. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer, pp 776–794, https://doi.org/10.1007/978-3-030-58621-8_45
Tian Y, Sun C, Poole B, et al (2020b) What makes for good views for contrastive learning? arXiv preprint https://doi.org/10.48550/arXiv.2005.10243
Tian Y, Henaff OJ, van den Oord A (2021) Divide and contrast: self-supervised learning from uncurated data. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10063–10074, https://doi.org/10.1109/ICCV48922.2021.00991
Tian Y, Xie L, Fang J, et al (2022) Beyond masking: demystifying token-based pre-training for vision transformers. arXiv preprint https://doi.org/10.48550/arXiv.2203.14313
Tian K, Jiang Y, Diao Q, et al (2023) Designing BERT for convolutional networks: sparse and hierarchical masked modeling. arXiv preprint https://doi.org/10.48550/arXiv.2301.03580
Tomasev N, Bica I, McWilliams B, et al (2022) Pushing the limits of self-supervised resnets: can we outperform supervised learning without labels on imagenet? arXiv preprint https://doi.org/10.48550/arXiv.2201.05119
Touvron H, Cord M, El-Nouby A, et al (2022) Three things everyone should know about vision transformers. In: European conference on computer vision, Springer, pp 497–515, https://doi.org/10.1007/978-3-031-20053-3_29
Tran M, Ly L, Hua BS, et al (2022) Ss-3dcapsnet: self-supervised 3d capsule networks for medical segmentation on less labeled data. In: 2022 IEEE 19th international symposium on biomedical imaging (ISBI), IEEE, pp 1–5, https://doi.org/10.1109/ISBI52829.2022.9761627
Träuble F, Creager E, Kilbertus N, et al (2021) On disentangled representations learned from correlated data. In: International conference on machine learning, PMLR, pp 10401–10412
Treneska S, Zdravevski E, Pires IM et al (2022) GAN-based image colorization for self-supervised visual feature learning. Sensors 22(4):1599. https://doi.org/10.3390/s22041599
Triantafillou E, Zhu T, Dumoulin V, et al (2020) Meta-dataset: A dataset of datasets for learning to learn from few examples. arXiv preprint https://doi.org/10.48550/arXiv.1903.03096
Tsai YHH, Wu Y, Salakhutdinov R, et al (2021) Self-supervised learning from a multi-view perspective. arXiv preprint https://doi.org/10.48550/arXiv.2006.05576
Tschannen M, Djolonga J, Rubenstein PK, et al (2020) On mutual information maximization for representation learning. arXiv preprint https://doi.org/10.48550/arXiv.1907.13625
Tukra S, Hoffman F, Chatfield K (2023) Improving visual representation learning through perceptual understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14486–14495, https://doi.org/10.1109/CVPR52729.2023.01392
Uelwer T, Robine J, Wagner SS et al (2025) A survey on self-supervised methods for visual representation learning. Mach Learn 114(4):1–56. https://doi.org/10.1007/s10994-024-06708-7
Van Berlo B, Hoey J, Wong A (2024) A survey of the impact of self-supervised pretraining for diagnostic tasks in medical x-ray, ct, mri, and ultrasound. BMC Med Imaging 24(1):79. https://doi.org/10.1186/s12880-024-01253-0

Van Engelen JE, Hoos HH (2020) A survey on semi-supervised learning. Mach Learn 109(2):373–440. https://doi.org/10.1007/s10994-019-05855-6

Van Horn G, Mac Aodha O, Song Y, et al (2018) The iNaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8769–8778, https://doi.org/10.1109/CVPR.2018.00914

Van Horn G, Cole E, Beery S, et al (2021) Benchmarking representation learning for natural world image collections. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12884–12893, https://doi.org/10.1109/CVPR46437.2021.01269

Veeling BS, Linmans J, Winkens J, et al (2018) Rotation equivariant cnns for digital pathology. In: Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11, Springer, pp 210–218, https://doi.org/10.1007/978-3-030-00934-2_24

Villarraga DF, Daziano RA (2025) Designing graph convolutional neural networks for discrete choice with network effects. arXiv preprint https://doi.org/10.48550/arXiv.2503.09786

Vo HQ, Yuan P, Yin Z, et al (2025) MIRAM: Masked image reconstruction across multiple scales for breast lesion risk prediction. arXiv preprint https://doi.org/10.48550/arXiv.2503.07157

Von Kügelgen J, Sharma Y, Gresele L et al (2021) Self-supervised learning with data augmentations provably isolates content from style. Adv Neural Inf Process Syst 34:16451–16467

Wagner D, Ferreira F, Stoll D, et al (2022) On the importance of hyperparameters and data augmentation for self-supervised learning. arXiv preprint https://doi.org/10.48550/arXiv.2207.07875

Wah C, Branson S, Welinder P, et al (2011) The Caltech-UCSD birds-200-2011 dataset

Wallace B, Hariharan B (2020) Extending and analyzing self-supervised learning across domains. In: European conference on computer vision, Springer, pp 717–734, https://doi.org/10.1007/978-3-030-58574-7_43

Wang F, Liu H (2021) Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2495–2504

Wang T, Isola P (2022) Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: arXiv preprint, https://doi.org/10.48550/arXiv.2005.10242

Wang X, Qi GJ (2023) Contrastive learning with stronger augmentations. IEEE Trans Pattern Anal Mach Intell 45(5):5549–5560. https://doi.org/10.1109/TPAMI.2022.3203630

Wang L, Yoon KJ (2022) Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans Pattern Anal Mach Intell 44(6):3048–3068. https://doi.org/10.1109/TPAMI.2021.3055564

Wang X, Peng Y, Lu L, et al (2017) Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2097–2106, https://doi.org/10.1109/CVPR.2017.369

Wang J, Jiao J, Liu YH (2020) Self-supervised video representation learning by pace prediction. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16, Springer, pp 504–521, https://doi.org/10.1007/978-3-030-58520-4_30

Wang X, Zhang R, Shen C, et al (2021) 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, https://doi.org/10.1109/CVPR46437.2021.00304

Wang Y, Albrecht CM, Braham NAA, et al (2022a) Self-supervised learning in remote sensing: a review. arXiv preprint https://doi.org/10.48550/arXiv.2206.13188

Wang W, Bao H, Dong L, et al (2022b) Image as a foreign language: beit pretraining for all vision and vision-language tasks. arXiv preprint https://doi.org/10.48550/arXiv.2208.10442

Wang X, Zhang Y, Zhang Z et al (2022c) GSC-MIM: Global semantic integrated self-distilled complementary masked image model for remote sensing images scene classification. Front Ecol Evol 10:1083801. https://doi.org/10.3389/fevo.2022.1083801

Wang WC, Ahn E, Feng D et al (2023a) A review of predictive and contrastive self-supervised learning for medical images. Mach Intell Res 20(4):483–513. https://doi.org/10.1007/s11633-022-1406-4

Wang X, Chen G, Qian G et al (2023b) Large-scale multi-modal pre-trained models: a comprehensive survey. Mach Intell Res. https://doi.org/10.1007/s11633-022-1410-8

Wang H, Fan J, Wang Y, et al (2023c) DropPos: Pre-training vision transformers by reconstructing dropped positions. arXiv preprint https://doi.org/10.48550/arXiv.2309.03576

Wang S, Gao J, Li Z, et al (2023d) A closer look at self-supervised lightweight vision transformers. In: arXiv preprint, https://doi.org/10.48550/arXiv.2205.14443

Wang X, Huang Y, Zeng D et al (2023e) CaCo: Both positive and negative samples are directly learnable via cooperative-adversarial contrastive learning. IEEE Trans Pattern Anal Mach Intell 45(9):10718–10730. https://doi.org/10.1109/TPAMI.2023.3262608

Wang H, Song K, Fan J, et al (2023f) Hard patches mining for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10375–10385, https://doi.org/10.1109/CVPR52729.2023.01000
Wang C, Jiang L, Wu X, et al (2024a) GroupContrast: Semantic-aware self-supervised representation learning for 3d understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 4917–4928, https://doi.org/10.1109/CVPR52733.2024.00470
Wang J, Xu Z, Yang X et al (2024b) Self-supervised multi-view clustering in computer vision: a survey. IET Comput Vis 18(6):709–734. https://doi.org/10.1049/cvi2.12299
Wang Y, Yan X, Hu C, et al (2024c) Generative and contrastive paradigms are complementary for graph self-supervised learning. In: 2024 IEEE 40th international conference on data engineering (ICDE), IEEE, pp 3364–3378, https://doi.org/10.1109/ICDE60146.2024.00234
Wanyan X, Seneviratne S, Shen S, et al (2024) Extending global-local view alignment for self-supervised learning with remote sensing imagery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) Workshops, pp 2443–2453, https://doi.org/10.1109/CVPRW63382.2024.00251
Wei J, Suriawinata A, Ren B, et al (2021) A petri dish for histopathology image analysis. In: Artificial intelligence in medicine: 19th international conference on artificial intelligence in medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings, Springer, pp 11–24, https://doi.org/10.1007/978-3-030-77211-6_2
Wei C, Fan H, Xie S, et al (2022a) Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14668–14678, https://doi.org/10.1109/CVPR52688.2022.01426
Wei Y, Hu H, Xie Z, et al (2022b) Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. arXiv preprint https://doi.org/10.48550/arXiv.2205.14141
Wei L, Xie L, Zhou W, et al (2022c) Mvp: Multimodality-guided visual pre-training. In: European conference on computer vision, Springer, pp 337–353
Wei Y, Hu H, Xie Z, et al (2023a) Improving clip fine-tuning performance. In: 2023 IEEE/CVF international conference on computer vision (ICCV), IEEE, pp 5416–5426, https://doi.org/10.1109/ICCV51070.2023.00501
Wei C, Mangalam K, Huang PY, et al (2023b) Diffusion models as masked autoencoders. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 16284–16294, https://doi.org/10.1109/ICCV51070.2023.01492
Wei Y, Gupta A, Morgado P (2024a) Towards latent masked image modeling for self-supervised visual representation learning. In: European conference on computer vision, Springer, pp 1–17, https://doi.org/10.1007/978-3-031-72933-1_1
Wei J, Liu Y, Huang X, et al (2024b) Self-supervised graph neural networks for enhanced feature extraction in heterogeneous information networks. In: 2024 5th International conference on machine learning and computer application (ICMLCA), IEEE, pp 272–276, https://doi.org/10.1109/ICMLCA63499.2024.10753970
Weinstein J, Collisson EA, Mills GB et al (2013) The cancer genome atlas pan-cancer analysis project. Nat Genet 45(10):1113–1120. https://doi.org/10.1038/ng.2764
Wen Z, Li Y (2021) Toward understanding the feature learning process of self-supervised contrastive learning. In: Proceedings of the 38th international conference on machine learning, proceedings of machine learning research, vol 139. PMLR, pp 11112–11122
Woo S, Debnath S, Hu R, et al (2023) Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16133–16142, https://doi.org/10.48550/arXiv.2301.00808
Wu Z, Xiong Y, Yu SX, et al (2018) Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3733–3742, https://doi.org/10.1109/CVPR.2018.00393
Wu B, Xu C, Dai X, et al (2020) Visual transformers: token-based image representation and processing for computer vision. arXiv preprint https://doi.org/10.48550/arXiv.2006.03677
Wu HH, Kao CC, Tang Q, et al (2021) Multi-task self-supervised pre-training for music classification. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 556–560, https://doi.org/10.1109/ICASSP39728.2021.9414405
Wu C, Wu F, Qi T, et al (2022) Noisytune: A little noise can help you finetune pretrained language models better. arXiv preprint https://doi.org/10.48550/arXiv.2202.12024
Wu L, Lin H, Tan C et al (2023) Self-supervised learning on graphs: contrastive, generative, or predictive. IEEE Trans Knowl Data Eng 35(4):4216–4235. https://doi.org/10.1109/TKDE.2021.3131584

Wu ZF, Mao C, Wang W, et al (2024) Structured model probing: empowering efficient transfer learning by structured regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16849–16858, https://doi.org/10.1109/CVPR52733.2024.01593
Wu H, Li B, Tian L et al (2025) Corev2: A unified contrastive and restorative self-supervised learning method with masked prompting tuning for surface defect inspection. IEEE Trans Instrum Meas. https://doi.org/10.1109/TIM.2025.3545193
Xiao Y (2024) Self-supervised learning in deep networks: a pathway to robust few-shot classification. In: 2024 International conference on image processing, computer vision and machine learning (ICICML), IEEE, pp 1818–1823, https://doi.org/10.1109/ICICML63543.2024.10957986
Xiao J, Hays J, Ehinger KA, et al (2010) Sun database: large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, pp 3485–3492, https://doi.org/10.1109/CVPR.2010.5539970
Xiao T, Wang X, Efros AA, et al (2021) What should not be contrastive in contrastive learning. arXiv preprint https://doi.org/10.48550/arXiv.2008.05659
Xiao Y, Jiang A, Liu C et al (2022) Semantic-aware automatic image colorization via unpaired cycle-consistent self-supervised network. Int J Intell Syst 37(2):1222–1238. https://doi.org/10.1002/int.22667
Xie Z, Lin Y, Yao Z, et al (2021a) Self-supervised learning with swin transformers. arXiv preprint https://doi.org/10.48550/arXiv.2105.04553
Xie Z, Lin Y, Zhang Z, et al (2021b) 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, https://doi.org/10.1109/CVPR46437.2021.01641
Xie Z, Zhang Z, Cao Y, et al (2022a) On data scaling in masked image modeling. arXiv preprint https://doi.org/10.48550/arXiv.2206.04664
Xie Z, Zhang Z, Cao Y, et al (2022b) SimMIM: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9653–9663, https://doi.org/10.1109/CVPR52688.2022.00943
Xie Y, Zhang J, Xia Y, et al (2022c) Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: European conference on computer vision. Springer Nature Switzerland, pp 558–575, https://doi.org/10.1007/978-3-031-19803-8_33
Xie Z, Geng Z, Hu J, et al (2023a) Revealing the dark secrets of masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14475–14485, https://doi.org/10.1109/CVPR52729.2023.01391
Xie J, Li W, Zhan X, et al (2023b) Masked frequency modeling for self-supervised visual pre-training. arXiv preprint https://doi.org/10.48550/arXiv.2206.07706
Xie Y, Xu Z, Zhang J et al (2023c) Self-supervised learning of graph neural networks: a unified review. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2022.3170559
Xie Z, Zhang Z, Cao Y, et al (2023d) On data scaling in masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10365–10374, https://doi.org/10.1109/CVPR52729.2023.00999
Xie Y, Gu L, Harada T et al (2024) Rethinking masked image modelling for medical image representation. Med Image Anal 98:103304. https://doi.org/10.1016/j.media.2024.103304
Xu J (2021) A review of self-supervised learning methods in the field of medical image analysis. Int J Image Graph Signal Process 13:33–46
Xu M, Yoon S, Wu C, et al (2023) PlantCLEF2023: A bigger training dataset contributes more than advanced pretraining methods for plant identification. In: CLEF (Working Notes), pp 2168–2180
Xu J, Lin Z, Zhou D, et al (2024) Dppmask: Masked image modeling with determinantal point processes. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2266–2276, https://doi.org/10.1109/WACV57701.2024.00226
Xu J, Chen Z, Yang S, et al (2025) Mentor: multi-level self-supervised learning for multimodal recommendation. In: Proceedings of the AAAI conference on artificial intelligence, pp 12908–12917, https://doi.org/10.1609/aaai.v39i12.33408
Xue H, Gao P, Li H, et al (2023) Stare at what you see: masked image modeling without reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 22732–22741, https://doi.org/10.1109/CVPR52729.2023.02177
Xue Y, Gan E, Ni J, et al (2024) Investigating the benefits of projection head for representation learning. arXiv preprint https://doi.org/10.48550/arXiv.2403.11391
Yamaguchi S, Fukuda T (2023) On the limitation of diffusion models for synthesizing training datasets. arXiv preprint https://doi.org/10.48550/arXiv.2311.13090
Yamaguchi S, Kanai S, Shioda T, et al (2021) Image enhanced rotation prediction for self-supervised learning. In: 2021 IEEE international conference on image processing (ICIP), pp 489–493, https://doi.org/10.1109/ICIP42928.2021.9506132

Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. Association for Computing Machinery, New York, NY, USA, GIS '10, p 270–279, https://doi.org/10.1145/1869790.1869829

Yang X, He X, Liang Y, et al (2020a) Transfer learning or self-supervised learning? A tale of two pretraining paradigms. arXiv preprint https://doi.org/10.48550/arXiv.2007.04234

Yang X, He X, Zhao J, et al (2020b) COVID-CT-dataset: a CT scan dataset about COVID-19. arXiv preprint https://doi.org/10.48550/arXiv.2003.13865

Yang J, Shi R, Ni B (2021) Medmnist classification decathlon: a lightweight automl benchmark for medical image analysis. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI), IEEE, pp 191–195, https://doi.org/10.1109/ISBI48211.2021.9434062

Yang H, Ding X, Wang J, et al (2022) SimCL: Simple contrastive learning for image classification. In: Proceedings of the 5th international conference on big data technologies, pp 273–278, https://doi.org/10.1145/3565291.3565335

Yang S, Ge Y, Yi K, et al (2023a) Rils: Masked visual reconstruction in language semantic space. arXiv preprint https://doi.org/10.48550/arXiv.2301.06958

Yang Y, Huang LK, Wei Y (2023b) Concept-wise fine-tuning matters in preventing negative transfer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 18753–18763, https://doi.org/10.1109/ICCV51070.2023.01719

Yang Y, Huang W, Wei Y, et al (2023c) Attentive mask clip. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2771–2781, https://doi.org/10.1109/ICCV51070.2023.00260

Yang H, Ren Z, Yuan H et al (2023d) Contrastive self-supervised representation learning without negative samples for multimodal human action recognition. Front Neurosci 17:1225312. https://doi.org/10.3389/fnins.2023.1225312

Yang J, Shi R, Wei D et al (2023e) Medmnist v2 - A large-scale lightweight benchmark for 2d and 3d biomedical image classification. Sci Data 10(1):41. https://doi.org/10.1038/s41597-022-01721-8

Yang L, Lin S, Zhang F, et al (2025) Efficient self-supervised continual learning with progressive task-correlated layer freezing. In: 2025 26th international symposium on quality electronic design (ISQED), pp 1–8, https://doi.org/10.1109/ISQED65160.2025.11014440

Yangyang Y, Xiang F (2019) A flower image classification algorithm based on saliency map and pcanet. J Commun Comput 15:1548–7709

Yao Y, Desai N, Palaniswami M (2022) Masked contrastive representation learning. arXiv preprint https://doi.org/10.48550/arXiv.2211.06012

Yao Y, Desai N, Palaniswami M (2023) MOMA: Distill from self-supervised teachers. arXiv preprint https://doi.org/10.48550/arXiv.2302.02089

Yerxa T, Feather J, Simoncelli E et al (2024) Contrastive-equivariant self-supervised learning improves alignment with primate visual area it. Adv Neural Inf Process Syst 37:96045–96070

Yi JSK, Seo M, Park J, et al (2022) PT4AL: Using self-supervised pretext tasks for active learning. In: Computer vision – ECCV 2022. Springer Nature Switzerland, pp 596–612, https://doi.org/10.1007/978-3-031-19809-0_34

Yi K, Ge Y, Li X, et al (2023) Masked image modeling with denoising contrast. arXiv preprint https://doi.org/10.48550/arXiv.2205.09616

Yin L, Ghosh R, Lin C et al (2023) Mapping smallholder cashew plantations to inform sustainable tree crop expansion in benin. Remote Sens Environ 295:113695. https://doi.org/10.1016/j.rse.2023.113695

Yu J, Li X, Koh JY, et al (2022) Vector-quantized image modeling with improved VQGAN. arXiv preprint https://doi.org/10.48550/arXiv.2110.04627

Yu J, Yin H, Xia X et al (2024) Self-supervised learning for recommender systems: a survey. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2023.3282907

Zaiem S, Parcollet T, Essid S et al (2022) Pretext tasks selection for multitask self-supervised audio representation learning. IEEE J Select Top Signal Process 16(6):1439–1453. https://doi.org/10.1109/JSTSP.2022.3195430

Zbontar J, Knoll F, Sriram A, et al (2019) fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv preprint https://doi.org/10.48550/arXiv.1811.08839

Zbontar J, Jing L, Misra I, et al (2021) Barlow twins: self-supervised learning via redundancy reduction. arXiv preprint https://doi.org/10.48550/arXiv.2103.03230

Zeng J, Xie P (2020) Contrastive self-supervised learning for graph classification. arXiv preprint https://doi.org/10.48550/arXiv.2009.05923

Zeng X, Abdullah N, Sumari P (2024) Self-supervised learning framework application for medical image analysis: a review and summary. BioMed Eng OnLine 23(1):107. https://doi.org/0.1186/s12938-024-01299-9

Zhai X, Kolesnikov A, Houlsby N, et al (2022a) Scaling vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12104–12113, https://doi.org/10.1109/CVPR52688.2022.01179
Zhai X, Wang X, Mustafa B, et al (2022b) Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 18123–18133
Zhang J, Ma K (2022) Rethinking the augmentation module in contrastive learning: learning hierarchical augmentation invariance with expanded views. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16650–16659
Zhang R, Isola P, Efros AA (2016) Colorful image colorization. In: European conference on computer vision, Springer, pp 649–666, https://doi.org/10.1007/978-3-319-46487-9_40
Zhang R, Isola P, Efros AA (2017) Split-brain autoencoders: unsupervised learning by cross-channel prediction. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 645–654, https://doi.org/10.1109/CVPR.2017.76
Zhang X, Chen J, Yuan J, et al (2022a) Cae v2: Context autoencoder with clip target. arXiv preprint https://doi.org/10.48550/arXiv.2211.09799
Zhang T, Gao P, Dong H et al (2022b) Consecutive pre-training: a knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain. Remote Sens 14(22):5675. https://doi.org/10.3390/rs14225675
Zhang Z, Girdhar R, Joulin A, et al (2021c) Self-supervised pretraining of 3d features on any point-cloud. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10252–10263, https://doi.org/10.1109/ICCV48922.2021.01009
Zhang Y, Hooi B, Hu D, et al (2021d) Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. arXiv preprint https://doi.org/10.48550/arXiv.2102.06605
Zhang B, Rajan R, Pineda L, et al (2021e) On the importance of hyperparameter optimization for model-based reinforcement learning. arXiv preprint https://doi.org/10.48550/arXiv.2102.13651
Zhang C, Zhang K, Pham TX, et al (2022) 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, https://doi.org/10.1109/CVPR52688.2022.01404
Zhang X, Chen J, Yuan J, et al (2023a) CAE v2: Context autoencoder with CLIP latent alignment. Trans Mach Learn Res
Zhang W, Deng L, Zhang L et al (2023b) A survey on negative transfer. IEEE/CAA J Automatica Sinica 10(2):305–329. https://doi.org/10.1109/JAS.2022.106004
Zhang C, Zheng H, Gu Y (2023c) Dive into the details of self-supervised learning for medical image analysis. Med Image Anal 89:102879. https://doi.org/10.1016/j.media.2023.102879
Zhang T, Zhuang Y, Chen H et al (2023d) Object-centric masked image modeling-based self-supervised pretraining for remote sensing object detection. IEEE J Select Top Appl Earth Observ Remote Sens 16:5013–5025. https://doi.org/10.1109/JSTARS.2023.3277588
Zhang C, Zhang C, Song J, et al (2023e) A survey on masked autoencoder for visual self-supervised learning. In: Proceedings of the thirty-second international joint conference on artificial intelligence, pp 6805–6813, https://doi.org/10.24963/ijcai.2023/762
Zhang Y, Zhu H, Yu S (2023f) Adaptive data augmentation for contrastive learning. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 1–5, https://doi.org/10.1109/ICASSP49357.2023.10095350
Zhang F, Guo X, Peng X, et al (2024a) MIMIC: Mask image pre-training with mix contrastive fine-tuning for facial expression recognition. arXiv preprint https://doi.org/10.48550/arXiv.2401.07245
Zhang T, Wei D, Zhu M et al (2024b) Self-supervised learning for medical image data with anatomy-oriented imaging planes. Med Image Anal 94:103151. https://doi.org/10.1016/j.media.2024.103151
Zhang Y, Yuan G, Wu H et al (2024c) Mae-gan: a self-supervised learning-based classification model for cigarette appearance defects. Appl Comput Intell 4(2):253–268. https://doi.org/10.3934/aci.2024015
Zhang D, Zhou F, Albu F, et al (2024d) Unleashing the power of self-supervised image denoising: a comprehensive review. arXiv preprint https://doi.org/10.48550/arXiv.2308.00247
Zhang Y, Zhu H, Song Z, et al (2024e) Geometric view of soft decorrelation in self-supervised learning. In: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining, pp 4338–4349, https://doi.org/10.1145/3637528.3671914
Zhang Y, Zhu R, Zhang S, et al (2024f) Feature augmentation for self-supervised contrastive learning: a closer look. In: 2024 International joint conference on neural networks (IJCNN), IEEE, pp 1–8, https://doi.org/10.1109/IJCNN60899.2024.10651086
Zhang G, Tan P, Fang X et al (2025a) A unified self-supervised learning framework for hyperspectral image classification. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3549277

Zhang J, Yang L, Mohammadabadi SMS et al (2025b) A survey on self-supervised learning: recent advances and open problems. Neurocomputing. https://doi.org/10.1016/j.neucom.2025.131409
Zhao X, Du T, Wang Y, et al (2023) ArCL: enhancing contrastive learning with augmentation-robust representations. arXiv preprint https://doi.org/10.48550/arXiv.2303.01092
Zhao Z, Alzubaidi L, Zhang J et al (2024) A comparison review of transfer learning and self-supervised learning: definitions, applications, advantages and limitations. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2023.122807
Zheng L, Guha N, Anderson BR, et al (2021) When does pretraining help? Assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In: Proceedings of the eighteenth international conference on artificial intelligence and law, pp 159–168, https://doi.org/10.1145/3462757.3466088
Zhong Y, Tang H, Chen J, et al (2022) Is self-supervised learning more robust than supervised learning? arXiv preprint arXiv:2206.05259
Zhou B, Lapedriza A, Khosla A et al (2018) Places: A 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009
Zhou Z, Sodha V, Rahman Siddiquee MM, et al (2019a) Models genesis: generic autodidactic models for 3d medical image analysis. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22, Springer, pp 384–393, https://doi.org/10.1007/978-3-030-32251-9_42
Zhou B, Zhao H, Puig X et al (2019b) Semantic understanding of scenes through the ADE20K dataset. Int J Comput Vis 127:302–321. https://doi.org/10.1007/s11263-018-1140-0
Zhou M, Li Z, Xie P (2021) Self-supervised regularization for text classification. Trans Assoc Comput Linguist 9:641–656. https://doi.org/10.1162/tacl_a_00389
Zhou Z, Qi L, Yang X, et al (2022a) Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 20856–20865, https://doi.org/10.1109/CVPR52688.2022.02019
Zhou J, Wei C, Wang H, et al (2022b) iBOT: Image BERT pre-training with online tokenizer. arXiv preprint https://doi.org/10.48550/arXiv.2111.07832
Zhou P, Zhou Y, Si C, et al (2022c) Mugs: A multi-granular self-supervised learning framework. arXiv preprint https://doi.org/10.48550/arXiv.2203.14415
Zhou L, Liu H, Bae J, et al (2023a) Self pre-training with masked autoencoders for medical image classification and segmentation. In: 2023 IEEE 20th international symposium on biomedical imaging (ISBI), IEEE, pp 1–6, https://doi.org/10.48550/arXiv.2203.05573
Zhou M, Liu X, Yi T et al (2023b) A superior image inpainting scheme using transformer-based self-supervised attention GAN model. Expert Syst Appl 233:120906. https://doi.org/10.1016/j.eswa.2023.120906
Zhou Q, Yu C, Luo H, et al (2023c) MimCo: Masked image modeling pre-training with contrastive teacher. arXiv preprint https://doi.org/10.48550/arXiv.2209.03063
Zhou X, Zeng Y, Gong Y (2023d) Learning to scale temperature in masked self-attention for image inpainting. arXiv preprint https://doi.org/10.48550/arXiv.2302.06130
Zhu J, Li Y, Hu Y et al (2020) Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. Med Image Anal 64:101746. https://doi.org/10.1016/j.media.2020.101746
Zhuang X, Li Y, Hu Y, et al (2019a) Self-supervised feature learning for 3d medical images by playing a rubik's cube. In: Medical image computing and computer assisted intervention – MICCAI 2019. Springer International Publishing, pp 420–428, https://doi.org/10.1007/978-3-030-32251-9_46
Zhuang C, Zhai AL, Yamins D (2019b) Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6001–6011, https://doi.org/10.1109/ICCV.2019.00610
Zhuang J, Wu L, Wang Q et al (2025) MiM: Mask in mask self-supervised pre-training for 3d medical image analysis. IEEE Trans Med Imaging. https://doi.org/10.1109/TMI.2025.3564382
Zini S, Gomez-Villa A, Buzzelli M, et al (2023) Planckian jitter: countering the color-crippling effects of color jitter on self-supervised training. arXiv preprint https://doi.org/10.48550/arXiv.2202.07993
Zong Y, Mac Aodha O, Hospedales TM (2024) Self-supervised multimodal learning: a survey. IEEE Trans Pattern Anal Mach Intell 47(7):5299–5318. https://doi.org/10.1109/TPAMI.2024.3429301