Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 93–101
DOI: 10.15439/2023F8371
ISSN 2300-5963 ACSIS, Vol. 35

Can Unlabelled Data Improve AI Applications? A Comparative Study on Self-Supervised Learning in Computer Vision.

Markus Bauer
Center for Scalable Data Analytics and Artificial Intelligence
Humboldtstraße 25, Leipzig, 04105 Germany
Email: bauer@wifa.uni-leipzig.de

Christoph Augenstein
Center for Scalable Data Analytics and Artificial Intelligence
Humboldtstraße 25, Leipzig, 04105 Germany
Email: augenstein@wifa.uni-leipzig.de

***Abstract*—Artificial Intelligence (AI) represents a highly investigated area of study at present and has already become an indispensable component within an extensive range of business models and applications. One major downside of current supervised AI approaches lies in the need of numerous annotated data points to train the models. Self-supervised learning (SSL) circumvents the need for annotation, by creating supervision signals such as labels from the data itself, rather than requiring experts for this task. Current approaches mainly include the use of generative methods such as autoencoders and joint embedding architectures to fulfil this task. Recent works present comparable results to supervised learning in downstream scenarios such as classification after SSL-pretraining. To achieve this, typically modifications are required to suit the approach for the exact downstream task. Yet, current review works haven't paid too much attention to the practical implications of using SSL. Thus, we investigated and implemented popular SSL approaches, suitable for downstream tasks such as classification, from an initial collection of more than 400 papers. We evaluate a selection of these approaches under real-world dataset conditions, and in direct comparison to the supervised learning scenario. We discuss SSL's potential to take up with supervised learning, as well as the influence of the right training methods. Furthermore, we also introduce future directions for SSL research, as well as current limitations in real-world applications.**

I. Introduction

Self-supervised learning (SSL) has recently gained massive attention as a promising new learning paradigm in the machine learning world. The main advantage over supervised learning lies in SSL's capability to reduce the amount of required prior work of data scientists by avoiding manual annotation. Recent work has shown that state-of-the-art results can be achieved in downstream tasks, when using SSL as a pretraining method. This includes classification scenarios, as well as clustering.

Especially for medical and industry applications, where annotations can only be created by highly trained, rarely available experts, this can possibly be a game-changer in bringing Artificial Intelligence (AI) to a wider mass of companies. Currently, however, SSL methods are typically evaluated on large image datasets such as ImageNet [1], which compare poorly to real-world data. The problem is that ImageNet and comparable datasets consist of well-balanced, well-curated and giant data collections. Real-world datasets, especially in the industry and medical sector, on the other hand, typically only feature a few 1000 to 10000 samples which contain systematic label noise, defective images and imbalanced data. Additionally, such datasets often contain obvious data properties and fine-granular ones, whereas the latter ones need to be represented by the SSL-extracted features. Popular SSL methods currently are only proven to work correctly if balanced data [2] and large batch sizes are used [3]. For datasets and tasks under real-world conditions, this is often unfeasible and thus, the applicability and required modifications of SSL remain unclear.

We thus reviewed the literature to find the most important approaches for SSL in computer vision and summarize their basic functions, as well as possible modifications. We focused on practical implications of the presented approaches and investigated possible application fields of different algorithmic groups and their modifications. Additionally, we compared the performance of SSL models and their supervised pendants. For this, we evaluated the models using two datasets composed of MNIST and Oxford-Flower+IIT-Pet data [4], [5], [6] with little dataset size and different complexities of problems to address, thus, with conditions that hold for real-world applications as well.

To retrieve an extensive and representative collection of SSL-approaches, various databases were queried, using the system as described in Bramer et al. [7]. We identified appropriate journals and conferences for the topic and selected appropriate search engines that contain these. Several well-recognized journals such as the "IEEE Transactions on Pattern Analysis and Machine Intelligence" cover the scope of our work. We selected 20 of the best-ranked journals that cover our scope and picked an initial database according to the best availability of these journals. Among other choices, such as Cabi, Inspec and Web of Science, EBSCO had the best availability and thus was selected. The EBSCO academic search premier contains more than 3000 peer reviewed journals. In addition, we used the IEEE database directly, as well as Google Scholar, to also include popular conference proceedings such as the MICCAI, CVPR and NeurIPS, as well as publishers like ACM Digital Library and less popular journals. EBSCO is also preferred because it offers the use of thesauri terms, which are useful for filtering the search results. From an initial collection of 1792 samples, we finally extracted 42 articles to be considered for our comparative study. The detailed literature selection process is shown in Fig. 1.

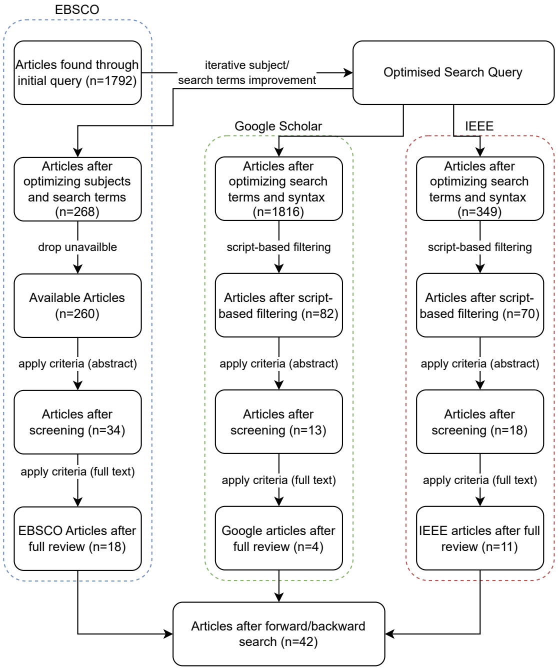

EBSCO
Articles found through initial query (n=1792)
iterative subject/ search terms improvement
Optimised Search Query
Google Scholar
IEEE
Articles after optimizing subjects and search terms (n=268)
Articles after optimizing search terms and syntax (n=1816)
Articles after optimizing search terms and syntax (n=349)
drop unavailble
script-based filtering
script-based filtering
Available Articles (n=260)
Articles after script-based filtering (n=82)
Articles after script-based filtering (n=70)
apply criteria (abstract)
apply criteria (abstract)
apply criteria (abstract)
Articles after screening (n=34)
Articles after screening (n=13)
Articles after screening (n=18)
apply criteria (full text)
apply criteria (full text)
apply criteria (full text)
EBSCO Articles after full review (n=18)
Google articles after full review (n=4)
IEEE articles after full review (n=11)
Articles after forward/backward search (n=42)

Fig. 1: Overview of the split search using EBSCO, IEEE Explore and Google Scholar.

II. Related Work

Various works have been proposed in the past that demonstrate impressively the value of SSL for clustering and classification tasks. Generally, two major approaches with five possible modification types can be found in the literature, among preliminary, less generalizable works. Fig. 2 provides an overview of SSL and its configuration possibilities.

Autoencoders (AEs) are the possibly most prominent architecture, and already gained attention for several years. The core principle is simple: Input data (i.e., images) is processed using a CNN encoder and a feature vector (:= embedding or latent) is created. Afterwards, the inverted CNN is used as a decoder to reconstruct the input. The supervision signal is then generated from the distance of the generated and input image. Thus, an optimization can be done, whereas reconstruction will improve with more meaningful latents that capture significant properties of the input data. A first application of this architecture was denoising, as proposed by Vincent et al. [8], followed by variants that make use of Gaussian inference [9] and additional classification tasks, called adversarial learning [10]. Even though AEs can achieve remarkable results for, e.g., clustering with smaller and simple image datasets (such as the MNIST dataset), they have known limitations when working with more complex data.

Thus, another class of SSL algorithms, joint embedding architectures (JEA), have gained attention recently. In JEA, the input will be processed using at least two CNNs (whereas weights may be shared), sent through a bottleneck to reduce dimensions, and then passed through a projection head, which is typically a fully connected network. As for the AE, backpropagation is the last step in JEA training. JEA uses a contrastive loss function [11] or an entropy-based function [12] that keeps the vectors of augmented and non-augmented image versions consistent. Very early works of Noroozi et al. [13] show that suitable SSL signals for JEA may be easy and intuitive to implement, e.g., by splitting or solving jigsaw puzzles. Even though such works didn't directly combine latents, and thus technically aren't JEAs, they share the idea and can be seen as close relatives of JEA. As the model needs to understand contextual information before solving the task, representative features will be learned as a side effect. Similar results can be achieved using image rotation [14], pseudo-classification [15] or combining multiple augmentations [11]. JEA's representative feature extraction capabilities can even improve if historical examples are used from a memory bank [16].

The major groups of AE and JEA can be divided further according to the following six modification types:

- Input variations: Input data is directly modified before being processed.
- Backbone variations: Other architectures than a single CNN are used.
- Latent variations: The features created by the bottleneck are processed in a parallel or preceding path to projection/decoding.
- Projection variations: A more sophisticated model than, e.g., an MLP is used as projection head.
- Temporal component: Time-dependent information is captured.

The following sections contain information about AE and JEA basic and modified approaches.

A. AE with Input Variations

The proposed basic principle of the AE as of Vincent et al. [8], as well as the ones of Kingma et al. and Makhzani et al. [9], [10], may be modified such that the input is preprocessed in a certain way, before feeding it to the AE. Such modifications may include augmentations, such as stretching, rotation and jitter, that will or will not be part of the target image. That way, the training can be guided to pay particular attention, or ignore certain properties of the input data, and thus to learn more robust features. This can be of value, e.g., in single-cell cytometric imaging, where AEs may be vulnerable to trivial features such as cell rotation [17]. Generally, an AE approach with input variations is seemingly mostly feasible, if obvious data properties should be ignored, as the typical application of this architecture often comes with little to no knowledge about important input data properties. This is not

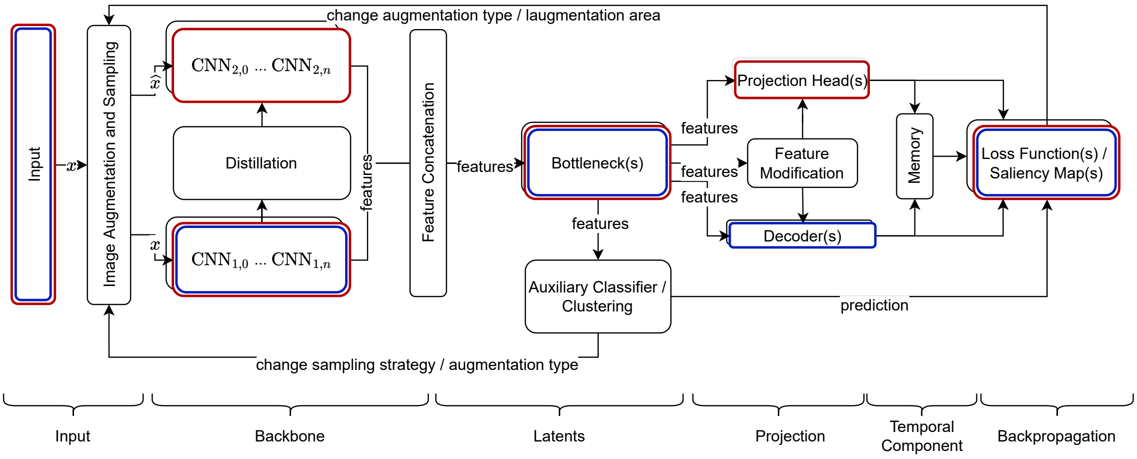

change augmentation type / laugmentation area
Input
x
Image Augmentation and Sampling
$\tilde{x}$
x
$\mathrm{CNN}_{2,0}$ ... $\mathrm{CNN}_{2,n}$
Distillation
$\mathrm{CNN}_{1,0}$ ... $\mathrm{CNN}_{1,n}$
features
Feature Concatenation
features
Bottleneck(s)
features
features
features
features
Projection Head(s)
Feature Modification
Decoder(s)
Memory
Loss Function(s) / Saliency Map(s)
Auxiliary Classifier / Clustering
prediction
change sampling strategy / augmentation type
Input
Backbone
Latents
Projection
Temporal Component
Backpropagation

Fig. 2: Overview of the discovered SSL training elements and their respective order. The default configurations for JEA (red) and AE (blue) are marked. Note that not all elements are necessarily used.

only limited to examples in the research field of life sciences. Another application in the industry context could, e.g., be the extraction of different features than build-part size in 3D-print classification. As the build-parts are printed layer after layer and hence their size differs throughout the print job, the most simple property is to learn exactly this varying size. This will, however, have almost no implications for, e.g., quality measurements. Using augmentations such as scaling and stretching, particular attention for anomalies can likely be created that enables extracting more meaningful latents that, e.g., correlate with the print quality.

B. AE with Backbone Variations

Another approach to increase the AE's performance lies in changing the encoder backbone from a simple CNN to a more sophisticated architecture. This includes using multiple backbone branches, or an encoder ensemble, whereas each model receives different input (i.e., augmented versions of the original). Furthermore, features from different model depths can be concatenated as the latent, to consider different morphological complexities in the vector representations. By modifying the backbone that way, a better weighting of global and local features may be achieved [18], as well as multiple views that carry different information may be considered for feature extraction, e.g., as in the case of analysing spectral bands in hyperspectral imaging (HSI) classification [19]. The concept can also be adapted to applications that leverage information, e.g., from audio and text data. Even though this variant may cause the trained models to be significantly more memory-intensive or slower, their feature extraction capabilities also may improve remarkably, when compared to a basic AE.

C. AE with Latent / Loss Variations

Similar to the approach of Makhzani et al. [10] the learning signal can be supported by not only formulating the cost functions as a reconstruction problem, but rather adding an auxiliary learning signal, such as the adversarial problem. Recent work shows that different learning signals than the adversarial discriminator, such as clustering, may be of value to assist the training process. They are included as a multi-loss cost function. Possible applications include the direct consideration of the extracted features clustering probabilities, e.g., in tasks where clustering of different cell lines is the goal [20]. Popular approaches to achieve better clustering capabilities include the use of Gaussian mixtures rather than a single Gaussian for the feature space [21]. The advantage of this modification is that it successfully extracts more robust features and thus is more suited against data distribution drift and noise. Typical application fields could also lie in life-science imaging, where artefacts are likely to arise and confuse trained models.

D. JEA with Backbone Variations

In contrast to the generative AE models, JEA enables one to train without the need for image generation. State-of-the-art results have recently been achieved in different downstream tasks of models pretrained with JEA. There are, however, applications, where the standard architecture needs to be modified analogous to the generative AE methods.

A first approach is to change the backbone or combine multiple architectures. The influence of different backbone models has been shown by Guerin et al. [22], who conclude that different architectures have different strengths/target domains and thus combining them helps to increase model performance. They use multiple pretrained architectures to perform clustering using JULE [23].

Another variant of backbone modifications is the use of Vision Transformers (ViTs) that could be shown to achieve comparable or even better results than CNNs that process images as a whole, while also, e.g., offering better properties when it comes to interpretability (c.f. [24]). They are thus a promising alternative to, e.g., residual networks [25]. In contrast to convolutional networks, ViTs show better properties regarding the weighting of local and global features, which can be advantageous in imaging domains that require context for each object, such as polarimetric synthetic aperture radar image classification [26]. For applications such as remote sensing imaging, where images are inherently very similar even though they may show different objects (e.g., a town house vs. a barn), ViTs also can be combined with knowledge distillation [27]. The imposed teacher vs. student asymmetry then improves fine-granular feature extraction capabilities. The distillation approach furthermore offers the possibility to reduce the network's parameters within the student model, and thus to decrease memory usage and processing times. JEAs backbone may also be modified, when multiple domains need to be unified, e.g., 2D and 3D images [28], or if special downstream tasks such as object detection are the training target [29].

E. JEA with Latent Variations

Directly modifying the latents may be of value to enable the model to, e.g., perform fine-grained clustering. This includes rearranging, normalization and regularization of the latents before processing them with the projection head and triggering the backpropagation. Additionally, created latents may also be used, to generate an auxiliary training signal that improves properties such as drift-resistance. This can be of interest, e.g., where slight changes between images have no meaning as in the case of scene classification, and thus the model needs to be robust against changes in poses, part configurations, as well as to relative motion between objects, and scene structures. Possible methods to implement such latent variations are spatial assembly [30] or the use of adversarial training for resistance against perturbations [31]. The use of an adversaria task also allows adapting JEA for image hashing problems [32].

Recent works as Masked Siamese Networks (MSN) or Joint Embedding Predictive Architectures (JEPA) randomly mask parts of the latents [33], [34], to simulate masking of the input images while maintaining asymmetry between, e.g., a student and teacher network. That way, the JEA scenario can be used with significantly less computation time, as no computation expensive augmentations need to be performed.

Modifications to, and further auxiliary processing of latents is advantageous where slight image changes have a significant impact on the data domain they refer to. This can, e.g., be the case in quality control, where small deviations from the standard can have a significant impact on product quality, in representation learning scenarios or in security applications.

F. JEA with Projection Variations

After latents have been extracted by the backbone, features typically are projected to a fixed-size vector, e.g., using fully connected layer and average pooling. Different methods allow for improving this projection compatible to later downstream applications such as clustering. For medical imaging, typically images need to be split into patches, and thus separate modules need to be created, that keep features consistent through the global image scope [35]. Furthermore, clustering capabilities of the extracted features may be improved, by assigning the features to cluster prototypes according to an optimal transport problem [36], which advantageously also avoids trivial solutions (mode collapse). Thus, using different projections heads, one can not only control the level of granularity in information extraction, but also adapt the model to special application scenarios, as in the case of histopathological whole-slide-imaging.

G. JEA with Temporal Component

In many scenarios, important information is time-dependent and thus can't be extracted from a single data point. Therefore, additional components need to be integrated, such that the model can capture the time-information. This can also enable a model to be suitable for streaming scenarios. Including temporal information can be done using distillation procedures [37], or memory bank approaches [38]. While distillation has the advantage of imposing asymmetry, which helps to find more robust features, memory banks can reduce the calculation efforts in terms of GPU-usage. Integrating a temporal component into the JEA learning setup further extends the possible application scope, e.g., to include quality monitoring, where slight data distribution changes are expected and need to be captured by the model to represent all data classes correctly [39]. This modification is, however, more of interest for time-series or video-based scenarios, than for imaging.

H. Preliminary Works

In addition to the presented groups and subgroups, further approaches have been discovered that show preliminary results or, as of now, fail to achieve state-of-the-art results on real-world data. They are presented for completeness here. Modifications to the default JEA scenario may also include recoupling of the model's output, e.g., by leveraging the Grad-Cam approach for sharpening the area of interest in the input image [40]. To ensure a higher probability of finding a global minimum during CNN parameter optimization, dropout is known as a useful measure, especially as it helps to avoid over-fitting. Thus, advanced dropout methods such as biologically inspired ones [41] may be a good choice to add to the JEA or AE model. The right choice of pre-text tasks also heavily influences the training. Thus, popular works often investigate how different augmentations contribute to the training results [42]. Other approaches show that methods typically used in JEA, such as the one of He et al. [43], are also of value for AEs. Examples include the mapping of AE features to dictionaries rather than simple keeping them for further processing [44] or combining AE and JEA architectures [45]. In special cases, selecting specific training variants, such as evolutionary [46] or sparse kernel network training, as well as solutions inspired by biological processes such as associative learning [47] may be of value to the JEA pipeline as well. A promising approach for future SSL directions lies in the usage of energy-based models (EBMs). These models may be capable of more fine-granular analysis of data distributions, and are thus in the focus of various visionary works, such as the one of Yann LeCun

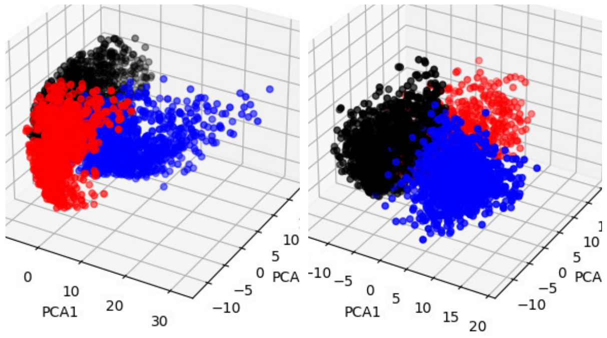
[48]. Restricted Boltzmann Machines are one very rudimentary implementation of such EBMs and have been studied in the literature, even though their capabilities in computer vision are very limited, as they model the data distribution, which is rarely feasible for images of a real-world size [49], [50]. Similar as with RBMs, self-organizing maps [51] are a popular choice, especially in life sciences and genetics, but have come unfashionable due to their poor scalability.

III. Qualitative and Quantitative Model Evaluation

To get a more profound understanding of the similarities and differences of SSL vs. supervised training, we analysed the observations and findings, and collected implications for SSL's applicability to real-world problems. We compared eleven different models, as depicted in Table I using three of the classes taken from the MNIST dataset, as well as various thousands images of cats, dogs, and flowers taken from the Oxford IIT-PET and Oxford IIT-Flower dataset. Note that each custom data set contains an obvious task and a more challenging task, as well as a slightly imbalanced data amount between the classes (e.g., more flowers than pets). For the MNIST data set, these are the separation of the digit four from eight and zero, and the separation of the similar digits zero and eight itself. Analogous, the Oxford-based data needs to be categorized into flowers and pets, as well as cats and dogs. The parameters for each training have been optimized in a grid search. For the qualitative analysis, the extracted latents were further reduced using a principal component analysis (PCA) and inspected. Similar results could be achieved for all models, except for the model using spatial transformations, as it struggled in finding a suitable solution. Fig. 3 shows the results of the highest accuracy models, as referred in Table II, for the MNIST dataset. It's evident that both models captured expressive filters and can successfully separate all the three digit types. The features extracted with the baseline model (SimCLR), however, seemed to use the latent space more efficiently, as they are denser. For the remaining models except the one with spatial transforms, distinguishable classes could be observed as well, while clusters were more entangled than in the presented examples.

TABLE I: Overview of implemented approaches and the subcategories. We implemented the particular methods as suggested in the referenced literature.

Approach	Group	Subgroup	References
AE Baseline	AE	-	[9]
AE + clustering	AE	Latent Variations	[21]
AE + ensemble	AE	Backbone Variations	[17], [19]
AE + input variations	AE	Input Variations	[17]
JEA baseline (SimCLR)	JEA	-	[11]
JEA + distillation	JEA	Backbone Variations	[27] w.o ViT
JEA + spatial transforms	JEA	Input Variations	[30]
JEPA	JEA	Latent Variations	[34]
MSN	JEA	Latent Variations	[34]
SwAV	JEA	Projection Variations	[52], [36]
Dino	JEA	Temporal Component	[53]



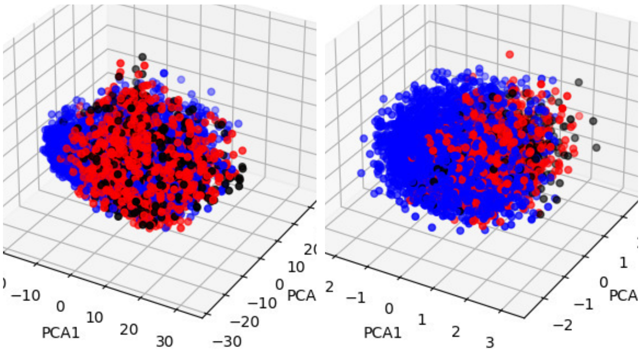
(a) AE with clustering.

(b) SimCLR.

Fig. 3: Features of a selection of MNIST samples (red – 0, black – 4, blue – 8), when converted to a three-dimensional space using the PCA.

For the Oxford datasets, qualitative results could be found to be worse than the ones for the MNIST dataset. This was not surprising, given the fact that the data contains far more complex objects. As shown in Fig. 4, the significant difference of flowers vs. pets could generally be recognized in the latent space's clusters, while the pets' features themselves yield only strongly entangled clusters. It's noteworthy to say that further compression through PCA possibly amplified the entangling, especially in the case of SimCLR. The remaining approaches showed similar behaviour for the Oxford datasets' qualitative analysis. The main difference was the amount of entangling.

From the qualitative analysis, three major observations could be made about SSL's capabilities. First, more complex datasets very likely require a higher number of samples, to guarantee finding a minimum on the error surface during training. Furthermore, fine-granular distinction of different data point classes, such as "cat" and "dog", may be challenging for both the AE and JEA setup. In addition, no significant difference between AE and SSL models could be observed in qualitative analysis, opposing to the results presented in further sections of this paper. The differences among the reviewed



(a) AE with clustering.

(b) SimCLR.

Fig. 4: Features of a selection of Oxford-Pets+Oxford-Flowers samples (red – cat, black – dog, blue – flower). In contrast to the MNIST data, the fine-grained task couldn't be solved.

modification types were investigated deeper, by evaluating model linear classification performance on the dataset's test cohorts, after pretraining with SSL. For the MNIST dataset, similar to the qualitative results, all approaches, except the spatial transforms, achieved results comparable to state-of-the-art models. The spatial transforms likely fail to achieve a good result, as the learned augmentations do not contribute to finding similar vector representations in the JEA setups. This effect is similar to the one observed by Li et al. [42] who conclude that only certain augmentation techniques are suited for their training task.

Overall, the AE with auxiliary (i.e., clustering) task performed best, in terms of absolute accuracy, with 99.87% (c.f. Table II). The JEA approaches achieved similarly good results, whereas SwAV even had the highest improvement compared to the supervised pendant of 1.37%. The results of the quantitative performance analysis thus matched the ones of the qualitative analysis. As the MNIST dataset could even be learned by linear machine learning models such as support vector machines, this finding can only be seen as proof of concept for the implemented approaches. For the Oxford data, more differences could be observed among different models. The previously top-performing AE with clustering only achieved a poor accuracy of 49.44% (c.f. Table II). Further investigation strongly suggested that the model failed to learn a meaningful distribution to generate data, thus the clustering loss was blocking the training, rather than helping, by creating confusing samples. Generally, the generative AE approaches performed worse than the JEA ones. This is likely because data complexity was too high for the small number of samples. As this is an unavoidable condition for numerous practical scenarios, generative AE approaches showed a systematic weakness here.

TABLE II: ΔAcc. (Accuracy – Supervised Accuracy) of different models on the MNIST and Oxford datasets, within a 5-fold cross-validation downstream test, when trained using SSL and in a supervised fashion.

Dataset	Approach	Acc. (%)	ΔAcc. (ppt.)	$\frac{t \cdot s^{-1}}{epoch}$
MNIST	AE baseline	99.46	0.03	29
	AE + clustering	**99.87**	1.29	36
	AE + ensemble	91.89	-6.87	87
	AE + input variations	99.31	1.21	153
	JEA + distillation	87.82	-8.65	47
	JEA + spatial transforms	50.44	-46.43	66
	SimCLR	99.72	1.36	58
	JEPA	91.45	-6.22	30
	MSN	93.21	-5.12	103
	SwAV	97.40	**1.37**	90
	Dino	90.36	-9.20	260
Oxford	AE Baseline	49.44	-20.26	17
	AE + clustering	55.82	-16.25	17
	AE + ensemble	56.27	-24.98	32
	AE + input variations	59.01	-21.1	46
	JEA + distillation	62.02	**-7.43**	21
	JEA + spatial transforms	34.32	-44.48	25
	SimCLR	**72.80**	-8.44	36
	JEPA	57.50	-24.19	17
	MSN	56.30	-25.39	58
	SwAV	55.40	-26.29	39
	Dino	61.90	-8.00	65

Among the JEA approaches, surprisingly, the baseline model performed best, with 72.80% accuracy. It's noteworthy to say that JEPA and MSN, however, took the shortest training time, as they already showed stable results after a few ten epochs. JEPA also had the fastest processing time per epoch. The distillation scenario showed the least difference to the supervised pendant, but similar to the generative AE approaches, the dataset size was possibly too small. The ViT-based Dino model likely suffers from the same problem. For the SwAV model, loss froze at an early stage of the training, at around epoch 50. This indicates that only a trivial solution was found in swapped prediction problem. Thus, training loss failed to converge to the global minimum. The effect could be compensated slightly by using a buffer to collect multiple batches, which increased the batches' variability, before solving the swapped prediction problem. The results were, however, inferior. The SimCLR approach, overall, could be found to work reasonably well, when compared to the supervised pendant. This also aligns with the findings of Zhong et al. [54].

The results on the Oxford datasets, opposing to the ones of the MNIST data, showed that systematic problems occurred for all the presented approaches, given a higher complexity of the data. Thus, the confusion matrices of the models we're examined, to see what the limitations are. Table III shows the result for the AE with clustering task, which was performing best on the MNIST dataset, and the JEA baseline. The AE with clustering model achieved a correct classification of flowers against the pets of 91%. For the more challenging part of distinguishing the pets themselves as cats, or respectively dogs, the model, however, performed only slightly better than a random classifier. The JEA Baseline provides better results with TP rates of 47, 77 and 95% for cats, dogs, and flowers. Similar to the AE, the less challenging task was solved remarkably well, while the dog vs. cat problem was characterized by a strong overfitting to the dog class.

TABLE III: Confusion matrix of the AE with latent variations and the JEA baseline on the Oxford-Flower+IIIT-Pet dataset.

	AE with clustering			SimCLR		
Pred True	Cat	Dog	Flower	Cat	Dog	Flower
Cat	0.23	**0.45**	0.31	**0.47**	0.24	0.29
Dog	0.16	**0.53**	0.31	0.07	**0.77**	0.17
Flower	0.03	0.06	**0.91**	0.01	0.04	**0.95**

As the training data only contained a few thousand images per class, the low data amount, in contrast to results created on the ImageNet dataset, seemed to be a limitation for all approaches. It's also noticeable that even more sophisticated JEA approaches generated worse solutions, as the results of the remaining models aligned with the ones of the clustering AE. An explanation for this may be that the dataset provided too few inputs to solve the optimization problem for a sufficient number of parameters, given the relatively small dataset size. This was especially the case for multi-loss training such as Dino, SwAV and JEA with distillation. The problem is likely caused by the non-contrastive loss-term. For example, in the clustering-based approaches (clustering AE and SwAV), the optimization of the clustering problem is only possible, if the batches are large enough at each step and have enough variability between the epochs. As this is not necessarily the case using a small dataset, the clustering-loss part will hinder training and eventually cause convergence to a suboptimal or trivial solution. Analogously, the second loss term of Dino and plain distillation do not contribute positively to the training. The AEs with ensemble and input variations showed improvements, when compared to the AE baseline, but couldn't keep up with the more capable JEA.

IV. Future Perspectives for SSL

From our results, we conclude that SSL can open new perspectives for future AI research. Even though the validated architectures failed to solve the most fine-grained task without specific modifications (i.e., separating pets and flowers of the Oxford data), various practical applications can be found in the literature that serve as a proof-of-concept for adapting SSL in industry and medical applications. In most of the cases identified in the literature, at least SSL-pretraining was helpful to extract more robust features or even such ones, that supervised learning was missing.

Future work should concentrate on facilitating the implementation of SSL in practice. Additionally, it should be investigated, how SSL can help in curating and understanding data, rather than simply using it as a tool to pretrain a model. Figure 5 shows such a scenario. After initially processing the unlabelled data, latents get annotated by an expert. To decrease the annotation burden, suggestions for annotation can be made by group selection of latents or labelling closest neighbours. That way, a larger dataset could be annotated using only a few examples. For uncertain latents, e.g., those that show similar logits for all pseudo classes, concepts such as active learning may be used for sampling. This way, experts could focus on conceptual work and result validation, rather than on searching for significant/anomalous data. Such a tool could bring major improvements to scenarios such as clinical applications, where initial annotation will likely result in subjective bias, or in scenarios where a-priori annotation is unfeasible due to the data set size.

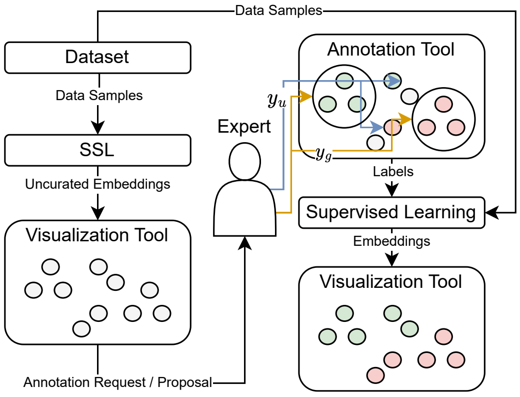


Fig. 5: SSL as a tool for data annotation.

V. Conclusions

In this work, we presented a comparative overview of current SSL methods. We conclude that SSL is a promising method to change the paradigms of machine learning, even though none of the approaches yet achieves identical or better performance than the supervised pendant on more complex datasets. The models solved less challenging tasks without problems, and showed promising initial results for more challenging tasks. Overall, they provided good baseline results that suggest SSL may be capable of achieving or surpassing performance of supervised training.

Regarding the training setups, we acknowledge that grid search may be a suboptimal choice to get optimal accuracies, especially as models with fewer hyperparameters benefit more from grid search than more complex ones due to search complexity. The performance results thus can only be seen as an indicator for certain characteristics of SSL approaches.

In addition to presenting the reviewed work, we performed an experimental validation using a selection of approaches. From the qualitative analysis, we conclude that all models generally capture important information from the data. The models, however, failed to solve the more challenging task. The latter finding is supported by the results of the performance analysis. Among all approaches, JEA methods outperformed generative (AE) ones. The SimCLR approach even showed a rudimentary solution to the challenging task of separating cats from dogs.

The biggest problem for all the models still seemed to be related to small dataset size. This may be the most significant weakness of SSL, as this condition typically can't be compensated. Additionally, many hyperparameters such as temperature [11], patch size (in case of using ViTs) or batch size need to be examined when using SSL (c.f. [3]). Especially the fact that approaches such as SimCLR, SwAV, MSN and JEPA require gigantic batch sizes of more than 1000 images, unnecessarily limits SSL applicability as multi-GPU clusters will be needed for calculation, with a data set of according size. In practice, this means a high technical burden to implement such a solution. Therefore, future research should focus on small-dataset SSL, that also works under real-world conditions, rather than focusing on ImageNet benchmarks. Additionally, more effort should be spent to understand structural differences between supervised and SSL models, and the exact effects leading to this behaviour.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2009.5206848

[2] M. Assran, R. Balestriero, Q. Duval, F. Bordes, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, and N. Ballas, "The hidden uniform cluster prior in self-supervised learning," *CoRR*, vol. abs/2210.07277, 2022.

[3] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A. G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum, "A cookbook of self-supervised learning," *CoRR*, vol. abs/2304.12210, 2023.

[4] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, 2012. [Online]. Available: http://dx.doi.org/10.1109/MSP.2012.2211477

[5] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1447–1454. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.42

[6] P. Omkar, M., V. Andrea, Z. Andrew, and J. C., V., "Cats and dogs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2012.6248092

[7] W. M. Bramer, G. B. D. Jonge, M. L. Rethlefsen, F. Mast, and J. Kleijnen, "A systematic approach to searching: an efficient and complete method to develop literature searches," *Journal of the Medical Library Association*, vol. 106, no. 4, Oct. 2018. [Online]. Available: http://dx.doi.org/10.5195/jmla.2018.283

[8] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, dec 2010. [Online]. Available: http://dx.doi.org/10.5555/1756006.1953039

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[10] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *CoRR*, vol. abs/1511.05644, 2015.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020.

[12] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *CoRR*, vol. abs/2103.03230, 2021.

[13] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," *CoRR*, vol. abs/1603.09246, 2016.

[14] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *CoRR*, vol. abs/1803.07728, 2018.

[15] Z. Wu, Y. Xiong, S. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance-level discrimination," *CoRR*, vol. abs/1805.01978, 2018.

[16] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," *CoRR*, vol. abs/1912.01991, 2019.

[17] L. Ternes, M. Dane, S. Gross, M. Labrie, G. Mills, J. Gray, L. Heiser, and Y. H. Chang, "A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis," *Communications Biology*, vol. 5, no. 1, 2022. [Online]. Available: http://dx.doi.org/10.1038/s42003-022-03218-x

[18] W. Xiong, L. Zhang, B. Du, and D. Tao, "Combining local and global: Rich and robust feature pooling for visual recognition," *Pattern Recognition*, vol. 62, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2016.08.006

[19] S. Zhang, M. Xu, J. Zhou, and S. Jia, "Unsupervised spatial-spectral cnn-based feature learning for hyperspectral image classification," *IEEE Transactions on Geoscience & Remote Sensing*, 2022. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2022.3153673

[20] C. Vununu, S.-H. Lee, and K.-R. Kwon, "A strictly unsupervised deep learning method for hep-2 cell image classification," *Sensors (14248220)*, vol. 20, no. 9, 2020. [Online]. Available: http://dx.doi.org/10.3390/s20092717

[21] V. Prasad, D. Das, and B. Bhowmick, "Variational clustering: Leveraging variational autoencoders for image clustering," *CoRR*, vol. abs/2005.046132, 2020.

[22] J. Guérin, S. Thiery, E. Nyiri, O. Gibaru, and B. Boots, "Combining pretrained cnn feature extractors to enhance clustering of complex natural images," *Neurocomputing*, vol. 423, 2021. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2020.10.068

[23] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.556

[24] X. Chen, C.-J. Hsieh, and B. Gong, "When vision transformers outperform resnets without pre-training or strong data augmentations," *CoRR*, vol. abs/2106.01548, 2021.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[26] H. Dong, L. Zhang, and B. Zou, "Exploring vision transformers for polarimetric sar image classification," *IEEE Transactions on Geoscience & Remote Sensing*, 2022. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2021.3137383

[27] X. Wang, J. Zhu, Z. Yan, Z. Zhang, Y. Zhang, Y. Chen, and H. Li, "Last: Label-free self-distillation contrastive learning with transformer architecture for remote sensing image scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022. [Online]. Available: http://dx.doi.org/10.1109/LGRS.2022.3185088

[28] W. Zhou, Y. Hou, K. Ouyang, and S. Zhou, "Exploring complementary information of self–supervised pretext tasks for unsupervised video pre–training," *IET Computer Vision (Wiley-Blackwell)*, vol. 16, no. 3, 2022. [Online]. Available: http://dx.doi.org/10.1049/cvi2.12084

[29] J. Ding, E. Xie, H. Xu, C. Jiang, Z. Li, P. Luo, and G.-S. Xia, "Deeply unsupervised patch re-identification for pre-training object detectors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2022.3164911

[30] Y. Li, S. Kan, J. Yuan, W. Cao, and Z. He, "Spatial assembly networks for image representation learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 871–13 880. [Online]. Available: http://dx.doi.org/10.1109/CVPR46437.2021.01366

[31] L. Fan, S. Liu, P.-Y. Chen, G. Zhang, and C. Gan, "When does contrastive learning preserve adversarial robustness from pretraining to finetuning?" *CoRR*, vol. abs/2111.01124, 2021.

[32] P. Feng and H. Zhang, "Self-supervised image hash retrieval based on adversarial distillation," in *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, 2022, pp. 732–737. [Online]. Available: http://dx.doi.org/10.1109/CACML55074.2022.00127

[33] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas, "Masked siamese networks for label-efficient learning," *CoRR*, vol. abs/2204.07141, 2022.

[34] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," *CoRR*, vol. abs/2301.08243, 2023.

[35] J. Yan, H. Chen, X. Li, and J. Yao, "Deep contrastive learning based tissue clustering for annotation-free histopathology image analysis," *Computerized Medical Imaging & Graphics*, vol. 97, pp. N.PAG–N.PAG, 2022. [Online]. Available: http://dx.doi.org/10.1016/j.compmedimag.2022.102053

[36] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *CoRR*, vol. abs/2006.09882, 2020.

[37] A. Gomez-Villa, B. Twardowski, L. Yu, A. D. Bagdanov, and J. van de Weijer, "Continually learning self-supervised representations with projected functional regularization," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 3866–3876. [Online]. Available: http://dx.doi.org/10.1109/CVPRW56347.2022.00432

[38] H. Kahng and S. B. Kim, "Self-supervised representation learning for wafer bin map defect pattern classification," *IEEE Transactions on Semiconductor Manufacturing*, vol. 34, no. 1, 2021. [Online]. Available: http://dx.doi.org/10.1109/TSM.2020.3038165

[39] W. Dai, M. Erdt, and A. Sourin, "Self-supervised pairing image clustering for automated quality control," *Visual Computer*, vol. 38, no. 4, 2022. [Online]. Available: http://dx.doi.org/10.1007/s00371-021-02137-y

[40] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, and T.-L. Liu, "Saga: Self-augmentation with guided attention for representation learning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3463–3467. [Online]. Available: http://dx.doi.org/10.1109/ICASSP43922.2022.9747302

[41] P. Yin, L. Qi, X. Xi, B. Zhang, and H. Qiao, "Nflb dropout: Improve generalization ability by dropping out the best -a biologically inspired adaptive dropout method for unsupervised learning," in *2016*

International Joint Conference on Neural Networks (IJCNN), 2016, pp. 1180–1186. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2016.7727331

[42] X. Li, X. Hu, X. Qi, L. Yu, W. Zhao, P.-A. Heng, and L. Xing, "Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, 2021. [Online]. Available: http://dx.doi.org/10.1109/TMI.2021.3075244

[43] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *CoRR*, vol. abs/1911.05722, 2019.

[44] K. Pang, K. Li, Y. Yang, H. Zhang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Generalising fine-grained sketch-based image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2019.00077

[45] J. Lu, L. Li, and C. Zhang, "Self-reinforcing unsupervised matching," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 44, no. 8, 2022. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2021.3061945

[46] X. Fang, Y. Cai, Z. Cai, X. Jiang, and Z. Chen, "Sparse feature learning of hyperspectral imagery via multiobjective-based extreme learning machine," *Sensors (14248220)*, vol. 20, no. 5, 2020. [Online]. Available: http://dx.doi.org/10.3390/s20051262

[47] J. Liu, M. Gong, and H. He, "Deep associative neural network for associative memory based on unsupervised representation learning," *Neural Networks*, vol. 113, 2019. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2019.01.004

[48] Y. LeCun, "A path towards autonomous machine intelligence," *under review*, 2022.

[49] J. Zhang, H. Wang, J. Chu, S. Huang, T. Li, and Q. Zhao, "Improved gaussian–bernoulli restricted boltzmann machine for learning discriminative representations," *Knowledge-Based Systems*, vol. 185, pp. N.PAG–N.PAG, 2019. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2019.104911

[50] B. Xiaojun and W. Haibo, "Contractive slab and spike convolutional deep boltzmann machine," *Neurocomputing*, vol. 290, 2018. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2018.02.048

[51] M. Sakkari, M. Hamdi, H. Elmannai, A. AlGarni, and M. Zaied, "Feature extraction-based deep self-organizing map," *Circuits, Systems & Signal Processing*, vol. 41, no. 5, 2022. [Online]. Available: http://dx.doi.org/10.1007/s00034-021-01914-3

[52] P. Goyal, Q. Duval, J. Reizenstein, M. Leavitt, M. Xu, B. Lefaudeux, M. Singh, V. Reis, M. Caron, P. Bojanowski, A. Joulin, and I. Misra, "VISSL," https://github.com/facebookresearch/vissl, 2021.

[53] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," *CoRR*, vol. abs/2112.13492, 2021.

[54] Y. Zhong, H. Tang, J. Chen, J. Peng, and Y.-X. Wang, "Is self-supervised learning more robust than supervised learning?" *CoRR*, vol. abs/2206.05259, 2022.