Training strategies with unlabeled and few labeled examples under 1-pixel attack by combining supervised and self-supervised learning

Gabriel B. Cavallari [1] **Moacir A. Ponti** [1 2]

Abstract

Self-supervised learning pre-training exhibited excellent performance on feature learning by using only unlabeled examples. Still, it is not clear how different self-supervised tasks perform under distinct image domains and there are still training issues to be tackled under scenarios of limited labeled data. We investigate two self-supervised tasks: rotation and Barlow Twins, on three distinct image domains, exploring a combination of supervised and self-supervised learning. Our motivation is to work on scenarios where the proportion of labeled data with respect to unlabeled data is small, as well as investigate the model's robustness to 1-pixel attacks. The models that combine supervised with self-supervised tasks can take advantage of the unlabeled data to improve the learned representation in terms of the linear discrimination, as well as allowing learning even under attack.

1. Introduction

Deep convolutional neural networks mainly follow the supervised learning paradigm, in which many input-output pairs are required for training (Razavian et al., 2014; Kornblith et al., 2019; Ponti et al., 2021). However, for a given task, manually labeled data are time-consuming and expensive to obtain, while unlabeled data is often widely available. Also, some applications involve patterns that are not present in standard large-scale benchmark datasets. Strategies using unlabeled and limited annotated data are paramount in this scenario (Cavallari et al., 2018; Dos Santos et al., 2020).

Self-supervised learning (Kolesnikov et al., 2019) is showing promising results using only unlabeled data, being able to learn features that are competitive with respect to supervised baselines (He et al., 2020; Chen et al., 2020; Grill et al., 2020; Caron et al., 2020). However, methods to improve training of self-supervised tasks are still to be investigated.

In this paper we aim at making the best use of small labeled training sets, with the additional challenge of not being able to use a validation set, and still leverage unlabeled data. For that we explore different training techniques with two self-supervised tasks: rotation-prediction (Gidaris et al., 2018) and Barlow Twins (Zbontar et al., 2021), and evaluate the learned representations in scenarios with unlabeled data, few labeled data and both unlabeled and few labeled data. Additionally, we evaluate model's robustness under 1-pixel attacks. We compare the methods using 3 image domains and show how unlabeled data with few labeled data can allow for more discriminative and robust representations compared to the supervised baselines.

2. Related Work

Self-supervised learning relies on pretext tasks formulated using unlabeled data. These models learn low-level features as good as via strong supervision (Asano et al., 2019) even under minimal data-learning (Shi et al., 2020). Semi-supervised learning (SSL) is a class of algorithms that learn considering both labeled and unlabeled data. Consistency regularization methods add auxiliary loss terms computed on the unlabeled data. The auxiliary loss terms can be considered as a regularizer. π-Model (Laine & Aila, 2017), Mean Teacher (Tarvainen & Valpola, 2017) and Virtual Adversarial Training (Miyato et al., 2018) that take advantage of consistency losses, among other works.

In terms of the semi-supervised training strategy and the use of the rotation-prediction as an auxiliary task, our work is most related to S4L (Zhai et al., 2019), SESEMI (Tran, 2019), (Hendrycks et al., 2019) and (Cavallari & Ponti, 2021). Zhai et al. (2019) train semi-supervised models with the rotation-prediction on the ImageNet dataset. Tran (2019) uses rotation-prediction task as an auxiliary loss term to train the model for SVHN, CIFAR-10 and CIFAR-100. Different from them, we tested also in Fashion-MNIST (a grayscale dataset) and the Malaria dataset, which is not angle oriented (unlike photographs that have angle bias). Cavallari & Ponti

[1] ICMC, Universidade de São Paulo, São Carlos-SP, Brazil [2] Mercado Livre, Brazil. Correspondence to: Gabriel Cavallari <gabriel.cavallari@usp.br>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

(2021) used only rotation and had unstable results due to the initialization sensitivity of the models, while we propose a new way to overcome this as well as exploring the recent Barlow Twins approach. Hendrycks et al. (2019) indicates that self-supervision increases the robustness to adversarial examples, label and input corruptions. We instead perform 1-pixel attacks, use a siamese network, and evaluate the feature representations and not directly the network model. Gidaris et al. (2019) and Su et al. (2020) also explore self-supervision as an auxiliary task but in a few-shot learning pipeline.

3. Method

We are interested in evaluating the feature representations resulting from using self-supervised pre-training methods, fine-tuning and semi-supervised alternatives for image classification problem on different domains, including natural and biomedical images. Our experiments consider a small labeled dataset D_l with N_l pairs of images and labels, and a much larger unlabeled dataset Du containing N_u images, from which five experiments were performed:
1. *Supervised training in limited annotated data*, using 1% or 5% of labeled data in relation to the total of unlabeled data. The annotated data are from the D_l set;
2. *Unsupervised training with RotNet or BarlowTwins*, which do not use labels, on the unlabeled D_u dataset;
3. *Semi-supervised training* by fine-tuning from the frozen weights of the models trained in step 2, using 1% or 5% of labeled data in relation to the total of unlabeled data;
4. *Unified semi-supervised* training that uses 1% or 5% of the labeled data in relation to the total of unlabeled data, that is trained simultaneously with the unlabeled data D_u through siamese architecture and unified cost function:

$$\mathcal{L}_{SS} = \lambda_l \cdot \ell_l(D_l) + \lambda_u \cdot \ell_u(D_u), \tag{1}$$

where both ℓ_l and ℓ_u optimize a cross-entropy loss function: the former with the traditional supervised paradigm, and the latter with a self-supervised task. Weights $\lambda_l, \lambda_u > 0$. It can be used under different self-supervised losses ℓ_u. The network has shared weights between the supervised task and the self-supervised task in the main architecture.
5. *Fine-tuning the semi-supervised* method from step 4, with 1%/5% of labels in relation to the total of unlabeled data.

We use a ResNet50 backbone, discarding the final classification layer and adding a Global Average Pooling layer, that outputs 2048 values. Then, we added a fully connected layer with ReLU with 128 dimensions, which outputs the final representation evaluated in the experiments.

For the semi-supervised method (see Figure 1), after a ResNet50 backbone, separate branches are used: one softmax layer for classification (classes' output), and another branch for self-supervision. For RotNet a softmax layer predicting rotation degrees 0, 90, 180 and 270. For Barlow Twins, two branches of the main network, that receive as input two altered views of the same image, are connected to a three-layer MLP. As in Zbontar et al. (2021), the first two layers have Batch Normalization and ReLu, and the third one does not, but we used 2048-d layers instead of 8192.

The Barlow Twins task uses the same data augmentation pipeline and transformations from the original paper. The rotation task only uses rotation. The labeled images are not modified with data augmentations.

3.1. Training procedure and learning rates

All models were trained from scratch. For all experiments $\lambda_l = \lambda_u = 1.0$. For experiments using labeled data, each training was performed using a random partition of labeled data, with the same labeled random datasets for supervised and semi-supervised training.

In the first exploratory experiments, we noticed that the convergence of self-supervised methods were highly sensitive to initialization and learning rates (LR). To overcome this, for each model we searched LR hyper-parameters: 0.01, 0.001 and 0.0001, with fixed exponential decay $e^{-0.01}$ after 5% of the total number of epochs. Thus, for every LR value 5 models are trained, each with a different weight initialization seed. Under the assumption that a small training set does not allow the 'luxury' of separating data for validation, the models used in testing are those 5 trained with the LR that achieves the lowest mean value of final training loss value. This way we select the most consistent LR for every model.

3.2. Datasets

We assess the performance of our experiments on: STL-10 (Coates et al., 2011), Fashion-MNIST (Xiao et al., 2017) and Malaria (Rajaraman et al., 2018). STL-10 is designed for semi-supervised and unsupervised feature learning, containing 96 × 96 RGB images with airplane, bird, car, cat, deer, dog, horse, monkey, ship and truck classes. Fashion-MNIST has 28 × 28 grayscale images with centered pieces of 10 classes of clothing and fashion accessories. Malaria contains instances of parasitized and uninfected cells from the thin blood smear slide images of segmented cells.

3.3. Experimental setup

Batch size of 32 was used in all experiments, except for Barlow Twins, since it benefits from larger batches. In this case we use batch size 200 for STL-10 and Fashion-MNIST, and size 120 for Malaria. For the STL-10 dataset, we used the original size of 96×96. For the Fashion-MNIST dataset, the images were upsized to 96×96. For the Malaria dataset, images were downsized to 128×128. The number of epochs was chosen empirically for each model, observing stabilization of the loss value between subsequent epochs.

The unified semi-supervised training has shared weights and receives minibatches with an equal amount of labeled and unlabeled images (balancing supervised and self-supervised tasks). At each epoch, the model sees all unlabeled N_u, while the labeled ones N_l are seen N_u/N_l times. Because $N_u > N_l$ and batches are balanced, in one epoch all unlabeled images are seen. Therefore, our network sees repeated labeled instances in an epoch. We compensate that by allowing more epochs for the supervised setting.

3.4. Evaluation

To assess the discriminative capacity of learned representations, we train a linear SVM to assess the accuracy as a proxy measure for linear separability (Mello & Ponti, 2018), on the features obtained by the 128-D layer just after the Global Average Pooling of ResNet50 backbone. Using the best 5 models (with different initialization seeds) trained in experiments 1 to 5 of as a feature extractor, we consider the following steps for evaluation:
– *1. Extract the training set representations*;
– *2. Train Linear SVM* (with $C = 1$) using the representations extracted in step 1;
– *3. Extract the test set representations*;
– *4. Test the SVM trained in step 2 on test set representations*.

We tested other C values for the SVM but it didn't change the results. The test set has never been seen during model or SVM training, and is only used to obtain the accuracies reported in Tables 1 and 2. The SVM was trained with 1,000 or 5,000 images (1% or 5%) for the STL-10 dataset; 600 or 3,000 images (1% or 5%) for Fashion-MNIST; 137 or 685 images (1% or 5%) for Malaria; and they are the same sets used in training the models when the model uses labeled data.

The 1-pixel attack case was produced assuming access to the training data. Thus, we attack only images contained in D_l. For each class we arbitrarily insert a white pixel at the same coordinate in all images. Since we do not use data augmentations for the labeled images, the location of the attacked single pixel is not affected.

4. Results

The STL-10 dataset originally consists of 100,000 images in the unlabeled set, 5,000 images in the training set and 8,000 images in the test set. When training the supervised models, we use either 1,000 or 5,000 images from the training set as labeled images (1% or 5% of images in relation to the total unsupervised set). When training the semi-supervised model, we use same fractions of labeled data but also use the whole 100,000 unlabeled images. When training the purely unsupervised (self-supervised) models we use only the whole set of 100,000 unlabeled images.

The Fashion-MNIST dataset originally consists of 60,000 images in the training set and 10,000 images in the test set. When training the supervised models, we use either 1% (600) or 5% (3,000) of labeled images. When training the semi-supervised model, we use the same fractions of labeled data but also use the whole 60,000 images as unlabeled data. When training the purely unsupervised (self-supervised) models we use only the whole set of 60,000 training images discarding the labels.

Malaria dataset originally consists of 27,558 labeled images in total. For our experiments we consider half of total images for the training set and the other half for the test set. When training the supervised baselines, we use either 1% (137) or 5% (685) of the training set. When training the semi-supervised model, we use the same fractions of labeled data but also use the whole training set of 13,779 images as unlabeled data. When training the purely unsupervised (self-supervised) models we use only the whole set of 13,779 unlabeled images.

The results for all datasets using regular data and attacked data are shown in Table 1 for the use of 1% labeled data, and 2 for 5% of labeled data. Mean and standard deviation of the 5 seeds are shown. Bold Values are the highest accuracies considering also the standard deviation superposition.

When using either 1% or 5% of available labeled data, the self-semi-supervised models excel in STL-10 and Fashion. In particular, the datasets under 1-pixel attack suffer from the use of supervised learning only, while self-semi-supervised training or pretraining allow significant improvement. This may indicate that the simultaneous training was able to help the model's robustness by guiding the cost function simultaneously in the self-supervised task, which uses unlabeled data, and in the common classification task, which used supervised data with noise.

When compared with the best results, we note that the fine-tuning of the semi-supervised model worsened the accuracy compared to the semi-supervised models without fine-tuning except for Fashion 5%. Interestingly, both self-supervised tasks alone achieved a more competitive accuracy than the supervised model when only 1% of labeled data is available.

It is noteworthy that Malaria with 5% of data was able to improve all methods in the non-attacked dataset, but the attacked one suffered significantly. Contrary to what usually happened with other datasets, fine-tuning the semi-supervised model brings an improvement to the accuracy in all scenarios, in both tasks and with both 1% and 5% of labeled data. Standard deviation for supervised models using 1% of data were large, indicating unstable training compared to other experiments in general.

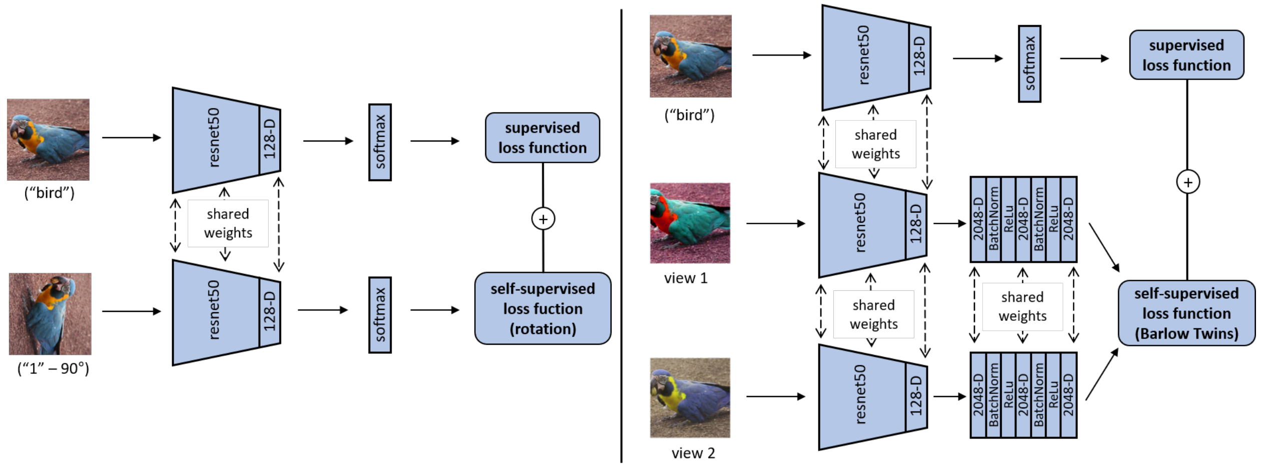

Figure 1. As shown on the left, when using the rotation prediction task, we have two separate softmax layers. One for the supervised classification task and the other for the rotation task. When using the Barlow Twins task, the other two branches of the main network are used by the Barlow Twins task, that receive as input two altered views of the same image, which will be passed through the backbone of the main architecture, and then connect to a three-layer MLP of size 2048, as shown on the right.

Table 1. SVM classification accuracies considering 1% of labeled data with regular and 1-pixel attacked (*) datasets.

Method	STL-10	STL-10*	Fashion	Fashion*	Malaria	Malaria*
Supervised	36.7 ± 0.4	35.4 ± 0.5	79.8 ± 0.8	75.7 ± 3.1	63.4 ± 10.8	66.7 ± 17.1
Rotation only	37.6 ± 1.0	-	70.9 ± 0.9	-	68.0 ± 1.4	-
Rotation only + fine-tuning	50.7 ± 1.0	51.1 ± 0.9	77.6 ± 0.9	77.0 ± 1.1	**88.4** ± 2.6	**81.9** ± 1.7
Rotation Semi	58.7 ± 1.3	56.5 ± 3.3	**83.1** ± 0.8	**78.2** ± 0.7	69.8 ± 2.8	66.3 ± 2.3
Rotation Semi + finetuning	55.3 ± 2.6	46.8 ± 4.4	81.3 ± 1.4	76.4 ± 3.1	72.8 ± 1.8	74.6 ± 3.0
Barlow Twins only	65.5 ± 0.7	-	71.7 ± 3.0	-	71.6 ± 6.9	-
Barlow Twins only + fine-tuning	69.5 ± 0.6	64.4 ± 0.4	81.1 ± 1.1	76.6 ± 1.4	67.5 ± 6.7	65.3 ± 4.0
Barlow Twins Semi	**70.3** ± 0.6	**70.4** ± 0.2	72.0 ± 3.7	73.6 ± 1.3	55.8 ± 5.2	55.8 ± 6.7
Barlow Twins Semi + finetuning	62.8 ± 2.4	61.5 ± 3.0	79.3 ± 2.5	71.8 ± 3.1	63.8 ± 5.2	61.9 ± 6.2

Table 2. SVM classification accuracies considering 5% of labeled data with regular and 1-pixel attacked (*) datasets.

Method	STL-10	STL-10*	Fashion	Fashion*	Malaria	Malaria*
Supervised	52.2 ± 0.8	47.4 ± 3.8	86.5 ± 0.4	76.8 ± 1.3	**94.1** ± 0.2	57.4 ± 2.4
Rotation only	41.5 ± 1.2	-	78.0 ± 1.6	-	71.8 ± 2.5	-
Rotation only + fine-tuning	63.9 ± 2.6	60.5 ± 3.3	87.5 ± 0.1	**83.0** ± 0.5	**93.9** ± 1.1	**90.3** ± 0.7
Rotation Semi	71.0 ± 0.6	60.2 ± 3.3	**88.3** ± 0.2	82.0 ± 1.1	87.2 ± 3.5	60.5 ± 5.2
Rotation Semi + finetuning	70.2 ± 0.3	58.0 ± 2.4	**88.0** ± 0.2	**83.0** ± 0.5	89.0 ± 3.1	71.6 ± 4.9
Barlow Twins only	73.1 ± 0.4	-	77.3 ± 2.1	-	75.0 ± 7.6	-
Barlow Twins only + fine-tuning	**77.8** ± 0.4	65.3 ± 1.8	86.7 ± 0.4	79.9 ± 0.8	89.0 ± 4.0	73.9 ± 9.6
Barlow Twins Semi	**77.4** ± 0.0	**76.1** ± 1.0	77.9 ± 1.0	78.6 ± 1.3	59.7 ± 2.7	61.5 ± 1.2
Barlow Twins Semi + finetuning	71.1 ± 0.6	64.5 ± 0.5	84.0 ± 1.6	72.7 ± 6.4	90.7 ± 2.7	73.2 ± 7.1

4.1. Discussion on the pretext tasks

Overall, the Barlow Twins performed better for the STL-10 dataset. This task uses *data augmentations* that can benefit scenarios where we have natural color images, which is perhaps the reason why we got good results for the STL-10. A better design of specific image pre-processing may yield better results for Fashion-MNIST and Malaria datasets.

The Rotation task performed better for the Fashion and Malaria datasets. The Semi-supervised models had the best performance in general, followed by the models that used some self-supervision + Fine-tuning task.

The biomedical image domain is not angle-oriented, thus the rotation task becomes harder. Nevertheless, significant improvement was found when incorporating unlabeled data. Also the attack had an even stronger impact on results due to the nature of the images, that contain patterns that are similar to the attack, degrading the results.

5. Conclusion

Obtaining large annotated datasets remains a limitation in training deep neural nets. Investigating different pre-training strategies that work under limited labeled data while leveraging unlabeled data is important. Combining self-supervision with supervised learning using rotation prediction and the Barlow Twins task is a good choice towards this objective. We also propose a way to overcome the sensitivity to initialization and learning-rate hyper-parameter by using the scarce training data only.

Our method goes beyond the use self-supervised tasks as pre-training, by simultaneously training supervised and self-supervised with a siamese architecture. Our semi-supervised model achieved overall the highest accuracy, followed by those using fine-tuning of a self-supervised pretext task. Not only we improve numerical results, but learn more discriminative spaces, as well as a more robust representation against 1-pixel attacks. Results show that the choice of the pretext task must take into account the nature of the dataset, and that a single task may not suit all applications.

Future work may investigate other types of auxiliary tasks in the context of semi-supervised learning, explore different weights for supervised/self-supervised losses, as well as test against other undesired scenarios of attack. In particular, we believe that handcrafting or learning pretext tasks for each dataset is a promising path.

Acknowledgments

The authors would like to thank FAPESP (grants #2019/07316-0 and #2019/02033-0) and CNPq (National Council of Technological and Scientific Development) grant 304266/2020-5.

References

Asano, Y., Rupprecht, C., and Vedaldi, A. A critical analysis of self-supervision, or what we can learn from a single image. In *International Conference on Learning Representations*, 2019.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. 2020.

Cavallari, G. B. and Ponti, M. A. Semi-supervised siamese network using self-supervision under scarce annotation improves class separability and robustness to attack. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 223–230. IEEE, 2021.

Cavallari, G. B., Ribeiro, L. S., and Ponti, M. A. Unsupervised representation learning using convolutional and stacked auto-encoders: a domain and cross-domain feature space analysis. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 440–446. IEEE, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conf. on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Coates, A., Ng, A., and Lee, H. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *AISTATS*, 2011.

Dos Santos, F. P., Zor, C., Kittler, J., and Ponti, M. A. Learning image features with fewer labels using a semi-supervised deep convolutional network. *Neural Networks*, 132:131–143, 2020.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., and Cord, M. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8059–8068, 2019.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, 2019.

Kolesnikov, A., Zhai, X., and Beyer, L. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1920–1929, 2019.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Mello, R. F. and Ponti, M. A. *Machine learning: a practical approach on the statistical learning theory*. Springer, 2018.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Ponti, M. A., dos Santos, F. P., Ribeiro, L. S., and Cavallari, G. B. Training deep networks from zero to hero: avoiding pitfalls and going beyond. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 9–16. IEEE, 2021.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568, 2018.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pp. 512–519. IEEE, 2014.

Shi, B., Hoffman, J., Saenko, K., Darrell, T., and Xu, H. Auxiliary task reweighting for minimum-data learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Su, J.-C., Maji, S., and Hariharan, B. When does self-supervision improve few-shot learning? In *European conference on computer vision*, pp. 645–666. Springer, 2020.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 1195–1204, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Tran, P. V. Exploring self-supervised regularization for supervised and semi-supervised learning. *arXiv preprint arXiv:1906.10343*, 2019.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Zhai, X., Oliver, A., Kolesnikov, A., and Beyer, L. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1476–1485, 2019.