Semi-supervised methods for CNN based classification of multispectral imagery

Manuel Bihler, Jiachen Zhou, and Michael Heizmann

Institute of Industrial Information Technology (IIIT), Karlsruhe Institute of Technology (KIT), Hertzstraße 16, 76187 Karlsruhe, Germany

Abstract Deep Convolutional neuronal networks, with their recent increase in performance, have become one of the standard techniques for RGB image classification. Due to a lack of large labeled datasets, this is not the case for multispectral image classification. To overcome this, we analyze the use of semi-supervised learning for the case of multispectral datasets. We use parameter reduction strategies to create small and efficient multispectral CNNs and combine these computationally efficient classifiers with semi-supervised learning methods. We choose the state-of-the-art semi-supervised methods MixMatch, ReMixMatch, FixMatch, and FlexMatch, to conduct experiments on the multispectral dataset EuroSAT. Additionally, we challenge this semi-supervised multispectral approach with a decreasing number of labeled images. We found that with only 15 labeled images per class, we can reach an accuracy above 80 %. If more labeled images are provided, the analyzed semi-supervised methods can even surpass basic supervised learning strategies.

Keywords Artificial intelligence, image processing, multispectral images, semi-supervised learning, CNN, consistency regularization, parameter reduction

1 Introduction

The use of deep convolutional neural networks for RGB image classification has led to a series of breakthroughs [1–4]. Extending convolutional neural networks to process multispectral imagery is becoming increasingly prevalent, especially in the field of characterization of materials, quality insurance in the food industry, or recycling of waste

materials [5]. In these fields, it is common to use multispectral (MS) data to separate materials based on their different spectral characteristics. While AI systems like CNNs show superior performance on large RGB datasets [1, 3, 4], the lack of large labeled multispectral datasets makes them difficult to employ in a multispectral setting. Compared to RGB images where there exist large publicly available datasets such as CIFAR-10 [6], and ImageNet [7], large labeled multispectral datasets are rare. In this work, we aim to improve the performance of CNNs on small unlabeled multispectral datasets by combining semi-supervised learning (SSL) methods with CNNs optimized for multispectral data (multispectral CNNs).

Semi-supervised learning provides a powerful tool to leverage unlabeled data and too largely alleviate the need for labeled data. This is particularly advantageous when collecting labeled data is expensive or time-consuming because expert knowledge or expensive machinery may be involved in the labeling process. This approach has shown impressive results in a wide variety of tasks, including facial expression recognition and natural language processing [8, 9].

To the best of our knowledge, the combination of SSL methods and multispectral CNNs is not discussed in previous work. We present a study on recently proposed state-of-the-art SSL methods in the context of classifying multispectral images. In this work, we show that modern SSL methods can be very effectively used to reduce the need for labeled data drastically. We also aim to make SLL methods more comprehensible for researchers outside the deep learning community. Therefore, in detail, we describe the methods used in the following section and then show results based on the EuroSAT dataset [10].

2 Semi-Supervised Methods

In image classification, semi-supervised learning (SSL) has proven to be a powerful paradigm for utilizing unlabeled data to mitigate the reliance on large labeled datasets. Compared with the results of previous SSL algorithms (π-Model [11], Mean teacher [12], Virtual Adversarial Training [13] and Pseudo-Label [14]), the four state-of-the-art SSL algorithms: MixMatch [15], ReMixMatch [16], FixMatch [17], and FlexMatch [18], all unify the current hybrid approaches for SSL. In this

section, we bring an overview of these four algorithms.

1. MixMatch: Unlike previous methods [11, 14], MixMatch introduces a single loss term unifying all three main semi-supervised approaches: entropy minimization [14, 19], consistency regularization [11,20] and generic regularization [21,22]. MixMatch utilizes a form of consistency regularization by using data augmentation for images. Two data augmentation methods are used subsequentially on both labeled and unlabeled images: first *random horizontal flip* and then *random crop*. Like Pseudo-Label [14], MixMatch applies multiple individual augmentations on an unlabeled image to create different instances, whose model predictions are then averaged to generate one pseudo-label for this unlabeled image. MixMatch uses a slightly changed version of the MixUp algorithm for regularization. Both labeled and unlabeled images and their corresponding labels are interpolated to generate mixed inputs and mixed labels.

2. ReMixMatch: To make MixMatch more data-efficient, two new techniques are introduced and directly integrated into MixMatch's framework: distribution alignment and augmentation anchoring. Distribution alignment maximizes the mutual information between model inputs and outputs so that unlabeled data is fully utilized to improve the model's performance. Distribution alignment encourages the marginal distribution of the model's predictions on unlabeled data to match the marginal distribution of the ground-truth labels. Recent work found that applying stronger forms of data augmentation can significantly improve the performance of consistency regularization [23]. Augmentation anchoring is added as a replacement for the consistency regularization in MixMatch. The basic idea is to use the model's prediction for a weakly augmented unlabeled image as the pseudo-label for many strongly augmented versions of the same image.

3. FixMatch: FixMatch is a significant simplification compared with MixMatch and ReMixMatch. Its simplification lies in combining only two main approaches to semi-supervised learning: consistency regularization and Pseudo-Label [14]. FixMatch first generates pseudo-labels on weakly augmented unlabeled images using their model predictions. For a given image, the pseudo-label is only retained if the model produces a high-confidence prediction. In other words, when the model assigns a probability to any class above the predefined threshold τ, the prediction is accepted, and the model output is then converted to a

one-hot pseudo label. Then, the model's prediction for a strongly augmented version of the same image is used to train the model against this pseudo-label.

4. FlexMatch: FixMatch uses a predefined constant threshold τ for all classes to select unlabeled data that contribute to the training, thus failing to consider different learning statuses and learning difficulties of different classes. To address this issue, Curriculum Pseudo Labeling (CPL) is introduced to utilize unlabeled data according to the model's learning status. The core of CPL is to adjust thresholds for different classes at each time step to feed the model with the fitting unlabeled data for the current learning status.

3 Results

In this section, we discuss our three main results. First, we present our classifier with a reduced number of parameters optimized for MS data and show the classification results on RGB and MS datasets, using supervised learning (SL). Secondly, we present the classification results using our classifier in combination with the above discussed SSL methods. Lastly, we show how the combination of MS data and SSL methods performs on datasets with a drastically decreased number of labeled images.

We use the datasets CIFAR-10 [24] and EuroSAT [10]. While CIFAR-10 is only used as a benchmarking dataset, EuroSAT is our main dataset for learning and testing the discussed strategies and methods. With 27,000 patches, EuroSAT is currently the largest labeled multispectral dataset for image patch classification. Additionally, it also contains the RGB bands, making it a perfect candidate for comparing RGB and MS learning strategies. Each multispectral image in the EuroSAT dataset consists of 13 channels, but only ten are relevant for identifying and monitoring land use classes and are used in our experiments. For the following experiments, we randomly sample 20 % and 10 % of labeled data from this dataset as validation and test sets respectively, while the remaining 18,900 labeled images are used as training data in either semi-supervised or fully supervised learning. We make sure that there is no overlap between these datasets.

3.1 Parameter Reduction

The success of deep neuronal networks like ResNet [25], or Wide ResNet [26], with their thousands of layers and millions of parameters, also lies in the availability of enormous datasets like CIFAR-10. In the case of multispectral imagery, where such datasets are lacking, very deep networks would easily overfit due to the extreme number of model parameters. Additionally, applying semi-supervised algorithms with deep CNNs as backbone classifiers can consume significant computational resources, making it a very costly and time-consuming combination of methods. To tackle this problem, we develop our own classifier optimized for the case of semi-supervised learning for multispectral imagery. This classifier is based on the Wide ResNet architecture and adopts parameter-reducing strategies presented in recent work on small and efficient CNNs, such as SqueezeNet [27] and MobileNet [28].

For further modification and evaluation, we choose the following Wide ResNet structures with fewer parameters while maintaining competitive accuracy according to the results in [26]: WRN-40-04, WRN-16-08, WRN-22-08 and WRN-28-10, where the first number depicts the *depth* and the second the widening factor k.

The structure of each residual block in the Wide ResNet consists of two 3x3 convolutional layers and hence is named B(3, 3), where B indicates the building block and (3, 3) the list of two kernel sizes of the convolutional layers. To decrease the number of parameters further, we additionally apply the microstructure from SqueezeNet [27] in every building block. Specifically, we replace all the 3x3 convolutional layers in each B(3, 3) building block with Fire Modules from SqueezeNet. In Figure 1 a sketch of the Fire module is depicted, and a detailed description of all variables used in the following is given in the caption. In each Fire Module, we set s_{1x1} equals to $0.125 \cdot C_{\text{In}}$, e_{1x1} equals to $0.75 \cdot C_{\text{Out}}$ and e_{3x3} equals to $0.25 \cdot C_{\text{Out}}$. The number of input and output channels of each 3x3 convolutional layer in the B(3, 3) block will be kept the same after replacement. The macro network structure of the original Wide ResNet will also be preserved. Hence, we call our network Wide ResNet with Fire Modules (WRN+FMs). It closely mimics the macro-architectural design of the Wide ResNet architecture while adapting the micro-architectural elements from the SqueezeNet to reduce network parameters.

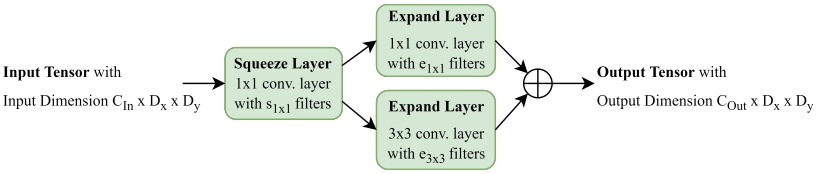

Figure 1: Fire Module structure as replacement for 3x3 convolutional layer.
C_{In}, C_{Out}: Number of input or output channels of the network block.
s_{1x1}: Number of output channels of the *Squeeze*-Layer.
e_{1x1}, e_{3x3}: Number of output channels of the 1x1 or 3x3 convolutional layer in the *Expand*-Layer, where $e_{1x1} + e_{3x3} = C_{\text{Out}}$.

We evaluate the new set of classifiers on two datasets, the RGB dataset CIFAR-10, and the multispectral dataset EuroSAT. In this section, we only use fully supervised learning to be able to compare our results with other SL benchmarks. For data augmentation, we do not use heavy data augmentation as proposed in semi-supervised learning algorithms and use only horizontal flips and random crops for images. Supervised training of Wide ResNet-28-10 (without FM) consumes too much training time and computing resources; therefore, we show results from literature [26, 29]. Our experimental results are shown in Table 1.

It can be concluded from Table1 that applying Fire Modules into the Wide ResNet structure brings benefits and also some expected downsides. With this parameter reduction strategy, the total number of network parameters can be significantly reduced, up to about 90% of the original network size. As a result, our WRN-28-10+FMs consists of only 2.42 million parameters and is 15 times smaller than the original WRN-28-10. Nevertheless, it achieves a classification accuracy of 96.19% on the EuroSAT MS dataset, only 0.41% less than the benchmark network SpectrumNet. From the results on EuroSAT in Table2, we find that WRN-28-10+FMs can achieve the best validation accuracy among our four new networks.

3.2 Semi-supervised Methods on MS data

We conduct experiments for the four selected SSL methods on the EuroSAT dataset using our classifier WRN-28-10+FMs and exhibit the re-

Table 1: Evaluation of different versions of Wide ResNet with and without Fire Modules on different datasets using fully supervised learning. The marked results are extracted from literature.

Dataset	Classifier	Parameter	Accuracy (%)
CIFAR-10 RGB	WRN-28-10	36.49 M	**95.83***
	WRN-28-10+FMs	2.40 M	92.51
	WRN-22-08	17.20 M	95.62*
	WRN-22-08+FMs	1.20 M	91.51
	WRN-16-08	11.00 M	95.19*
	WRN-16-08+FMs	0.86 M	90.79
	WRN-40-04	8.90 M	95.03*
	WRN-40-04+FMs	0.57 M	90.25
	SpectrumNet	0.72 M	92.29
EuroSAT Multispectral	WRN-28-10+FMs	2.42 M	96.19
	WRN-22-08+FMs	1.21 M	95.76
	WRN-16-08+FMs	0.87 M	94.89
	WRN-40-04+FMs	0.58 M	94.25
	SpectrumNet (Benchmark)	0.73 M	**96.60***

sults in Table 2. For semi-supervised learning, the number of labels for RGB and MS imagery is limited to 165 per class, i.e., the total number of labeled images for training is 1,650. This represents 6% of the entire dataset. The number of unlabeled images is set to 4,000 for both RGB and MS datasets to create a more realistic setting, as collecting high-dimensional MS images is more expensive and time-consuming. For comparison against supervised learning, we also conduct experiments using four different numbers of labeled images: (i) 5,650 to mimic the semi-supervised setting with the same number of samples: 4,000 unlabeled and 1,650 labeled images; (ii) 1,650 labeled images to simulate the same number of labeled images; (iii) 850 images and; (iv) 18,900 images to test the (unfair) lower and upper limit of supervised learning.

Table 2 show that all four SSL methods can still help our network achieve comparative classification accuracy, even though only limited labeled data is used. As expected, the supervised approach with the full amount of labeled images performs the best, with 96.56%. However, if the total number of labels is reduced to 5,650, the supervised method is outperformed by the semi-supervised method ReMixMatch

by 0.69%, although only 165 labeled images are used per class. One reason for this advantage of ReMixMatch lies in the utilization of strong data augmentation applied on both labeled and unlabeled images, which improves the performance of consistency regularization and helps the network achieve better robustness to noisy data. In general, MS images are expected to result in greater classification accuracy than RGB images in theory, given the additional information that is present in the spectral bands and increases the separation between classes. Except for MixMatch, all methods meet our expectations and perform better under MS conditions by 1.37% on average.

Table 2: Results of different semi-supervised learning methods on EuroSAT RGB and MS dataset using our WRN-28-10+FMs as classifier. Supervised learning with 850 and 18,900 images are not comparable with the SSL methods, they show the upper and lower limit of the methods for benchmarking purpose.

Dataset	SSL Methods	Accuracy (%)
EuroSAT RGB	MixMatch	94.64
	ReMixMatch	94.78
	FixMatch	88.28
	FlexMatch	92.91
EuroSAT MS	MixMatch	91.61
	ReMixMatch	**95.18**
	FixMatch	90.20
	FlexMatch	94.71
	SL with 850 images	68.65
	SL with 1,650 images (same number of labels)	78.33
	SL with 5,650 images (same number of samples)	94.49
	SL with 18,900 images	96.56

3.3 Limited number of labeled images

In this section, we drastically decrease the number of labeled images to test the limit of the discussed semi-supervised methods. The number of labeled MS images is decreased to 15, 30, 85 images per class, which represents only 0.5 %, 1% and, 3% of the entire dataset, while keeping the total number of unlabeled images the same with 4,000. This procedure is similar to other benchmarks in the literature [15–18].

The results from Figure 2 show that the classification performance of the network becomes better with an increasing number of labeled samples used in training. Among all the SSL methods, ReMixMatch consistently outperforms the other methods. FlexMatch follows ReMixMatch and proves to be the second best. The reason for this trend can be concluded as following: on the one hand, distribution alignment in ReMixMatch not only minimizes the entropy of pseudo labels for unlabeled data like all the other SSL methods do but also maximizes the mutual information between model inputs and outputs to incorporate unlabeled data for better model performance. On the other hand, a rotation loss [30] is directly included in the ReMixMatch loss term. Comparing SSL and SL for the case of 85 images per class drastically shows the power of semi-supervised learning. The SL approach with 850 images can only reach a classification accuracy of 68.65%, while the best SSL method reaches 95.07%.

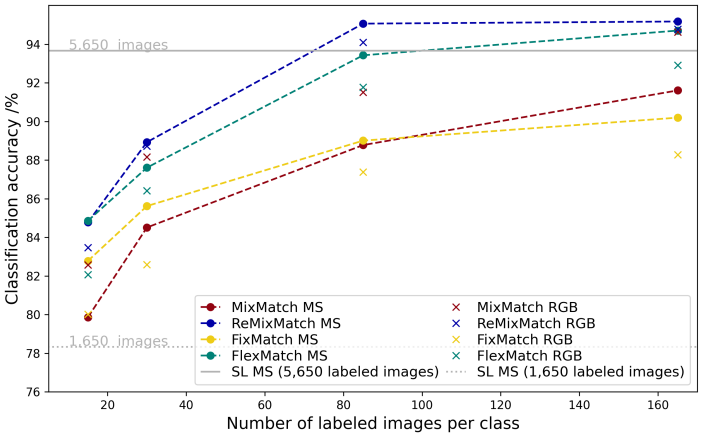

Figure 2: Results for the four SSL methods with a limited number of labeled images. For the SSL methods, 4,000 unlabeled images are available in addition to the depicted number of labeled images. For supervised learning, a gray solid/dashed line is shown for the case of the same number of samples (5,650 images) and the same number of labeled images (1,650 images), respectively.

4 Conclusions and Outlook

By adjusting the macro size of the Wide ResNet architecture and changing the micro-structure according to the SqueezeNet architecture, we obtain a small and efficient network with up to 15 times fewer parameters. We show that this network can compete with other popular networks on RGB datasets and can also be effectively trained on much smaller multispectral datasets. Based on the increased computational speed, it can be combined with modern SSL methods for RGB and multispectral datasets. To the best of our knowledge, the combination of SSL methods compressed CNNs, and multispectral datasets, have not been discussed in previous work. This work proves that using 85 images per class, state-of-the-Art SSL methods reach similar or even higher accuracies than supervised learning, depending on the augmentation strategies of the supervised approach. By decreasing the number of labeled images to 15 per class, the power of semi-supervised learning becomes even more prevalent, with 84.78% compared to SL 78.33% (1,650 images). Our results show that the newest SSL method in our comparison ReMixMatch outperforms the other methods not only for RGB but also for multispectral data.

These results show that SSL can be applied to MS data, and expensive labeling can be reduced dramatically. However, more research is needed to improve the number of augmentation strategies for multispectral data. Data augmentation plays a vital role in semi-supervised learning. There are still only a few specialized data augmentations available for multispectral channels compared with RGB channels. In future work, we are interested in investigating data augmentation methods for multispectral imagery according to the characteristics of different channels. We expect that the shown methods can increase the total number of available labeled datasets, which would benefit the whole research community in the field of image classification.

Acknowledgement

The project ASKIVIT is funded by the German Federal Ministry of Food and Agriculture (BMEL) through the Fachagentur Nachwachsende Rohstoffe e. V. under the funding reference 2220HV048A.

References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
2. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
3. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.
4. D. Matthew Zeiler and F. Rob, "Visualizing and understanding convolutional neural networks." ECCV, 2014.
5. S. L. Rabano, M. K. Cabatuan, E. Sybingco, E. P. Dadios, and E. J. Calilung, "Common garbage classification using mobilenet," in *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*. IEEE, 2018, pp. 1–4.
6. A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
7. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
8. S. Roy and A. Etemad, "Analysis of semi-supervised methods for facial expression recognition," *arXiv preprint arXiv:2208.00544*, 2022.
9. C. Liu, M. Zhang, Z. Fu, P. Hou, and Y. Li, "Flitext: A faster and lighter semi-supervised text classification with convolution networks," *arXiv preprint arXiv:2110.11869*, 2021.
10. P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.
11. S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.
12. A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

13. T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.
14. D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.
15. D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.
16. D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019.
17. K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.
18. B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021.
19. Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.
20. M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.
21. I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," 2018.
22. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
23. Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.
24. A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

25. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

26. S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

27. F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

28. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

29. J. J. Senecal, J. W. Sheppard, and J. A. Shaw, "Efficient convolutional neural networks for multi-spectral image classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

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