Self-supervised Pretraining for Hyperspectral Classification of Fruit Ripeness

Leon Amadeus Varga*, Hannah Frank*, and Andreas Zell

University of Tuebingen, Cognitive Systems
Sand 1, 72076 Tuebingen
* : shared contribution

Abstract The ripeness of fruit can be measured in a non-destructive way using hyperspectral imaging (HSI) and deep learning methods. However, the lack of labeled data samples limits hyperspectral image classification. This work explores self-supervised learning (SSL) as pretraining for HSI classification of fruit ripeness. Three state-of-the-art SSL methods, *SimCLR*, *SimSiam*, and *Barlow Twins* are implemented, and augmentation techniques for HSI are developed. A 3D-2D hybrid convolutional network is proposed to support the pretraining procedure. This model is evaluated against a *ResNet-18* and a *HS-CNN*. The pretraining is evaluated on the fruit ripeness prediction task using the proposed second version of the *DeepHS* fruit data set. Besides comparing the classification performance of the pretrained models to only supervised training, the influence of the model architecture and size, pretraining method, and augmentations for SSL is investigated. This work shows that it is possible to transfer the ideas of SSL to HSI. It is possible to extract essential features in an unsupervised manner via this pretraining. Pretraining stabilizes classifier training and improves the classifier performance. Further, it can partially compensate for the need for large labeled data sets in HSI classification.

Keywords Self-supervised learning, pretraining, hyperspectral imaging, HSI classification, fruit ripeness

1 Introduction

Knowing the ripeness of fruit is of great interest in the food industry. Especially exotic fruit, like avocados, kiwis, or papayas, are harvested when still unripe, kept in storage rooms, and are often shipped for weeks from far away. In addition, those kinds of exotic fruit often have a relatively high price. A reliable estimation of the fruit's ripeness state is required.

For this, usually, chemical and physical indicators like the sugar content and fruit flesh firmness are employed, all of which are obtained by destructive measurement.

It is also possible to predict the ripeness of fruit using hyperspectral imaging (HSI) [1, 2], which is non-destructive and therefore has become increasingly popular in recent years. Current work shows that combining HSI and deep learning can improve those predictions even further [3–5].

However, deep neural networks are usually trained in a supervised manner. Obtaining the actual ripeness state of a fruit still comes with destroying it, making the labeling process tedious and labeled samples scarce. Training networks on small training sets can be challenging, and overfitting becomes likely. Therefore, it is desirable to also use unlabeled fruit recordings that can be obtained without much effort.

Self-supervised learning (SSL) methods have produced astonishing results in computer vision [6–8] and may be applied for pretraining in this particular case of hyperspectral image classification to stabilize the training and potentially improve the network's predictions.

2 Experiments

2.1 Data Set

This work extended the already publicly available hyperspectral fruit data set, *DeepHS* [5], by additional recordings of avocados, kiwis, mangos, persimmon, and papayas. We used the same measurement setup and proceeding described by Varga et al. [5]. Each fruit was recorded by the *Specim FX 10* with 224 bands (398 nm - 1004 nm) and the *Corning microHSI 410 Vis-NIR Hyperspectral Sensor* with 249 bands (408 nm - 901

nm). Labels (firmness, sugar level, and overall ripeness) were obtained by destructive measurement.

The resulting *DeepHS v2* data set consists of 4671 recordings in total, 1018 labeled. Only the labeled subset was used for classification, while for self-supervised pretraining, also the unlabeled samples were used.

2.2 Models

Varga et al. [5] already proposed the *HS-CNN* network, a small convolutional neural network specialized for HSI data and the application for fruit ripeness classification.

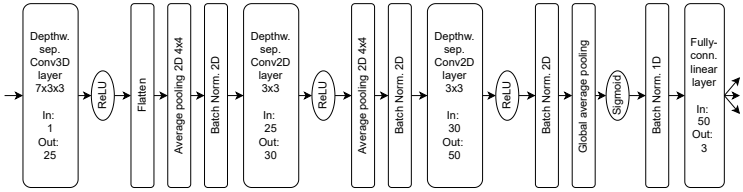


Figure 1: Architecture of the 3D-2D hybrid model.

We suggest a slightly modified variant, a 3D-2D hybrid model, using a 3D convolution instead of a 2D convolution in the first layer – inspired by *HybridSN* [9]. Its architecture is shown in Fig. 1. The backbone consists of a 3D convolutional layer for spectral-spatial feature learning and two 2D convolutional layers for more abstract spatial feature learning. Finally, a fully-connected layer operating on the spectral dimension is used for actual classification. With the hybrid version, we obtained a larger model ($\approx 20\times$ as many parameters than the baseline).

Additionally, we evaluated our methods using a *ResNet* architecture [10], which is also commonly employed for self-supervised learning (e.g., [6–8]) but has significantly more parameters compared to the other two models.

2.3 Self-supervised Pretraining

The model was pretrained using one of the three SSL methods: *SimCLR* [6], *SimSiam* [7], *Barlow Twins* [8].

All employ a siamese network architecture [11] where each branch is built by the encoder, the convolutional part of the classifier model, followed by a projection head. For the latter, we used a MLP with two layers. A *ReLU* non-linearity and batch normalization [12] was applied for each layer. The input dimension was 50 (for the baseline or hybrid model, and 512 for the *ResNet-18*), the hidden dimension was 16, and the embedding dimension was eight. For *SimSiam*, we used an additional prediction MLP, consisting of a single linear layer with input and output dimension of eight. The temperature parameter for *SimSiam* was chosen to be $\tau = 0.1$. For *Barlow Twins*, a weighting factor $\lambda = 0.01$ was used.

A critical component of SSL are the data augmentations. We evaluated 21 augmentation techniques, including four basic image transformations (rotating, flipping, cropping, random noise), two more specific ones (wavelength-dependent noise and pixel-wise intensity scaling), 13 augmentations that modify parts of the hyperspectral cube (i.e., drop or blur specific pixels, channels, or an entire sub-cube [13]), as well as two mixing augmentations (inspired by *MixUp* [14] and *ScaleMix* [15]).

Based on the ablation studies (see Sec. 4), only a subset of the augmentations (random rotations with probability 50%, random cropping with probability 30%, modification of the hyperspectral cube, and mixing with probability 20%) was actually used for pretraining.

The networks were optimized with SGD [16] with a weight decay of 10^{-4}, a momentum of 0.9, and a learning rate of 10^{-2}, decayed with the cosine decay schedule without restart [17]. We trained for 80 epochs with an effective batch size of 32.

2.4 Evaluation

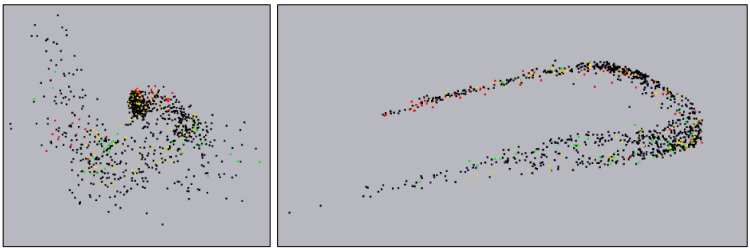
For the evaluation of self-supervised pretraining, the produced embeddings were considered. They were evaluated qualitatively (based on 3D visualizations) and quantitatively (based on the k-Nearest-Neighbor accuracy). For the visualization, the feature values of the embedding were plotted in three-dimensional space, after applying PCA. k-Nearest-Neighbor (k-NN) classification [18] was employed for the embedded labeled samples, using $k = 5$, the cosine distance and leave-one-out cross-validation (see, e.g., [7, 19]).

Additionally, we measured the performance for classification without and with pretraining. For the pretrained model, first, the fully-connected part was trained on top of the pretrained backbone, and then all model weights were further fine-tuned on the classification task (e.g., [6–8]). Without pretraining, the randomly initialized model was trained using settings similar to Varga et al. [5].

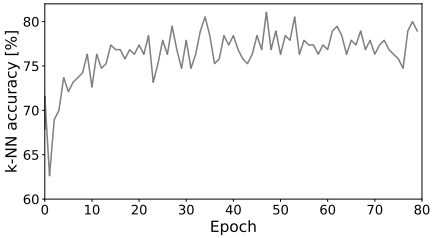
After the supervised training, the model was evaluated on the test set. Test time augmentations [20] were applied with probability 50%.

Using five different seeds each, we conducted experiments for all possible combinations of fruit types, cameras, and categories.

3 Results



(a) Embedding, before (left) and after pretraining (right).



(b) k-NN accuracy.

Figure 2: (a) 3D visualization of the embedding before and after pretraining via *Barlow Twins* – coloring by ripeness levels: unripe (green), ripe (yellow), overripe (red) and unlabeled (black). (b) k-NN accuracy on the ripeness levels of the labeled samples (train and validation set) during pretraining with *SimCLR*. For the hybrid model and the avocados, recorded by the *Specim* camera.

To evaluate the pretraining per se, we visualized the embeddings in 3D and monitored the k-NN accuracy during pretraining (see Fig. 2).

The spatial arrangement in the 3D space correlates with the ripeness level; samples of the same ripeness level are brought closer together. This fits the development of the k-NN accuracy, which increases as pretraining advances and finally converges towards 80%. This shows that pretraining can extract meaningful features and find useful representations for the data, without using label information.

Table 1: Classification accuracies (median, IQR) for regular classifier training versus *SimCLR* pretraining plus fine-tuning, for the *HS-CNN* (baseline) and hybrid model. One example for the five different fruit: Avocado (ripeness, *Specim*), kiwi (sugar, *Specim*), mango (firmness, *Specim*), kaki (sugar, *Specim*), papaya (ripeness, *Corning*), and over all fruit, categories and camera types. Highest accuracies in **bold**.

		Avocado	Kiwi	Mango	Kaki	Papaya	Overall
Baseline	Without pretraining	83.3% (±4.2%)	65.2% (±4.3%)	50.0% (±33.3%)	50.0% (±4.3%)	77.8% (±11.1%)	55.6% (±32.2%)
	With pretraining	87.5% (±0.0%)	73.9% (±8.7%)	50.0% (±8.3%)	**66.7%** (±8.7%)	**88.9%** (±0.0%)	**58.3%** (±32.2%)
Hybrid	Without pretraining	75.0% (±4.2%)	73.9% (±13.0%)	50.0% (±33.0%)	58.3% (±13.0%)	**88.9%** (±11.1%)	54.2% (±33.3%)
	With pretraining	**91.7%** (±4.2%)	**78.3%** (±4.3%)	50.0% (±16.7%)	58.3% (±4.3%)	**88.9%** (±11.1%)	**58.3%** (±36.1%)

Further, the pretrained model was evaluated on the downstream classification task. Especially, classification performance with pretraining and additional fine-tuning was compared to classification without pretraining.

We present the classification accuracy per fruit in Tab. 1.

The pretraining led, for all examples, to a performance improvement. We achieved an overall classification accuracy of 58.3%. Comparing the baseline model initially designed for pure classification to our newly proposed hybrid model with pretraining, overall, we could observe an improvement of approx. 3% in classification accuracy. For some fruit, it could be increased by more than 10%. Where this was not the case, the IQR was reduced, indicating that pretraining increased stability.

Further, experiments, visible in Fig. 3, show that pretraining even could compensate for the need for large amounts of labeled samples.

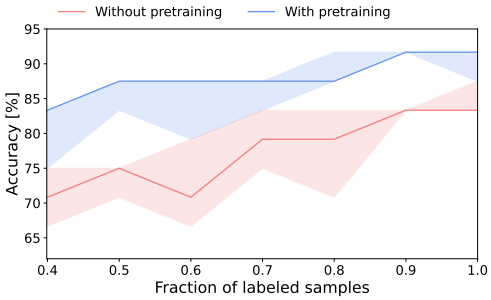


Figure 3: Classification accuracy (median and IQR) versus fraction of labeled samples used for classifier training for the baseline model with default classifier training (red) and hybrid model with pretraining (via *SimCLR*) plus fine-tuning (blue). Example: Avocado, *Specim* camera, ripeness classification.

4 Ablation Study

4.1 Classifier Model

For each of the three models, the classification accuracy with and without pretraining is visualized in Fig. 4.

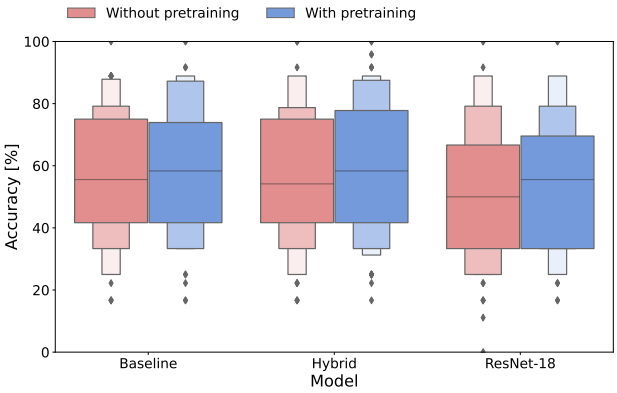


Figure 4: Classification accuracies for the *HS-CNN* baseline, hybrid and *ResNet-18* model, without pretraining (red) and with pretraining via *SimCLR* (blue).

For classification without pretraining, the *HS-CNN* performs best among all three models (55.6% accuracy). With pretraining, the performance can be improved only by a small amount, probably due to the affected backbone extracting only spatial and no spectral features.

The hybrid model, with 54.2% accuracy, performs slightly worse for classification without pretraining than the baseline, possibly due to overfitting. However, more importantly, with pretraining, the accuracy improved by a larger amount – reaching equal accuracy (58.3%) and indicating that a more powerful backbone makes pretraining more effective for the hybrid variant.

The *ResNet-18* performs worse than the other two models without and with pretraining. Again, this is probably due to overfitting and spatial feature extraction. However, it has the most significant improvement (more than 5%) by pretraining.

Overall, pretraining improved the classification accuracy relative to classification without pretraining. This improvement is more significant for larger models. We claim that pretraining can prevent overfitting and enables the training of larger models.

4.2 Self-supervised Pretraining Method

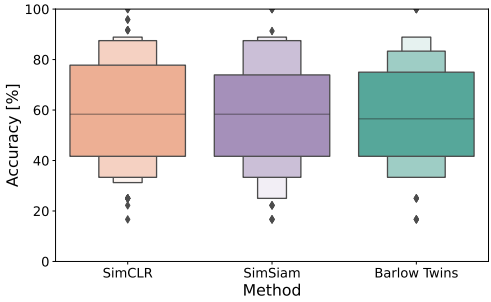


Figure 5: Classification accuracies for pretraining via *SimCLR*, *SimSiam*, *Barlow Twins* using the hybrid model. Over all fruit, categories and both cameras.

Secondly, we compare the three pretraining methods employed [6–8].

Although their approaches are very different, the classification performance is rather similar (visualized in Fig. 5). Overall, *SimCLR* per-

formed best, slightly better than *SimSiam*, which both have a median classification accuracy of 58.3%. *Barlow Twins* obtains only 56%.

4.3 Augmentations

Further, we evaluated the influence of the 21 proposed data augmentation techniques, by grouping them and using only one group for pretraining, respectively. Fig. 6 shows the resulting classification accuracies for the avocado fruit as a representative example.

The basic augmentations (rotating, flipping, cropping, and cutting) showed the highest accuracy ($> 80\%$) and therefore seemed to be most important. The pixel augmentations, like the modification of edge pixels and dropping random or consecutive pixels, were also helpful for pretraining. On the other hand, dropping multiple consecutive channels led to the worst classification accuracy ($< 70\%$). Also, dropping or blurring visible color channels decreased performance.

In general, distorting the spectrum resulted in low classification ac-

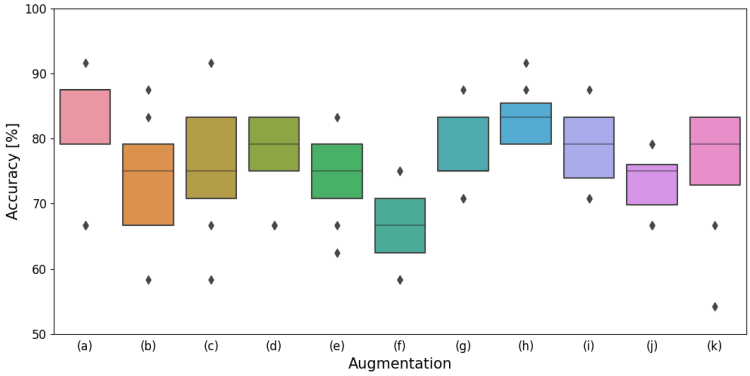


Figure 6: Classification accuracies for self-supervised pretraining (via *SimCLR*) using only the group of (a) basic augmentations, (b) noise augmentations, (c) augmentations that blur or drop random pixels, (d) drop consecutive pixels, (e) blur or drop random channels, (f) drop consecutive channels, (g) drop a subcube, (h) blur or drop edge pixels, (i) blur or drop edge channels, (j) blur or drop visible color information channels, and (k) mixing augmentations. Over all three SSL methods. Example: Avocado, *Specim*, ripeness classification.

curacy. We found that, for hyperspectral image data, introducing noise systematically instead of entirely random is more valuable.

5 Conclusion

In this work, the hyperspectral data set of ripening fruit was extended by two new measurement series and three new fruit types.

Further, we show that it is possible to transfer the ideas of SSL to hyperspectral data. SSL pretraining extracts essential features in an unsupervised manner and allows using larger models. It can stabilize classifier training and improves the classification accuracy in some situations. Therefore, pretraining can partially compensate for the need for large labeled data sets in HSI classification.

Fig. 7 shows the improvements achieved using SSL pretraining for the ripeness classification for the five different fruit. The classification accuracy could be boosted by more than 10% for the avocados and also for the kiwis. For mangos, kakis, and papayas, the classification itself is not stable, but for the papayas as well as overall, pretraining could reduce the variability. Summarizing, the pretraining allows a more reliable ripeness classification for specific exotic fruit.

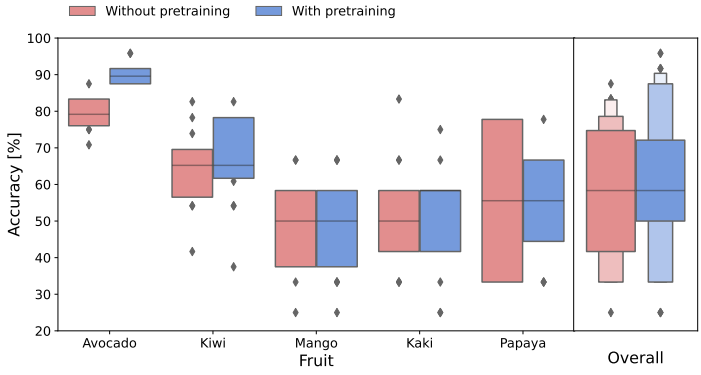


Figure 7: Classification accuracies for the baseline model without pretraining (red) versus the hybrid model with *SimCLR* pretraining (blue). For the *Specim* camera and the five different fruit (avocado, kiwi, mango, kaki, papaya), classified by all three categories (ripeness, firmness and sugar content).

References

1. O. O. Olarewaju, I. Bertling, and L. S. Magwaza, "Non-destructive evaluation of avocado fruit maturity using near infrared spectroscopy and PLS regression models," *Scientia Horticulturae*, vol. 199, pp. 229–236, 2016.
2. J. Pinto, H. Rueda-Chacon, and H. Arguello, "Classification of hass avocado (persea americana mill) in terms of its ripening via hyperspectral images," *TecnoLogicas*, vol. 22, pp. 111 – 130, 08 2019.
3. Z. Gao, Y. Shao, G. Xuan, Y. Wang, Y. Liu, and X. Han, "Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning," *Artificial Intelligence in Agriculture*, vol. 4, pp. 31–38, 2020.
4. C. Manliguez and J. Y. Chiang, "Multimodal deep learning and visible-light and hyperspectral imaging for fruit maturity estimation," *Sensors*, vol. 21, p. 1288, 02 2021.
5. L. A. Varga, J. Makowski, and A. Zell, "Measuring the ripeness of fruit with hyperspectral imaging and deep learning," in *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*. IEEE, 2021, pp. 1–8.
6. T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607.
7. X. Chen and K. He, "Exploring simple siamese representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 15 750–15 758.
8. J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 12 310–12 320.
9. S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-d-2-d CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 2, pp. 277–281, 2020.
10. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778.

11. J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.

12. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 448–456.

13. J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Hyperspectral image classification using random occlusion data augmentation," *IEEE Geosci. Remote. Sens. Lett.*, vol. 16, no. 11, pp. 1751–1755, 2019.

14. H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

15. X. Wang, H. Fan, Y. Tian, D. Kihara, and X. Chen, "On the importance of asymmetry for siamese representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 16 549–16 558.

16. J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, 1952.

17. I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *CoRR*, vol. abs/1608.03983, 2016.

18. E. Fix and J. L. Hodges, "Discriminatory analysis. Nonparametric discrimination; consistency properties," USAF School of Aviation Medicine, Randolph Field, Texas, Tech. Rep., 1951.

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

20. A. G. Howard, "Some improvements on deep convolutional neural network based image classification," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.