Visual Harvest: Self-supervised learning for lettuce growth analysis

Andrei Simion-Constantinescu[1,a] and Joaquin Vanschoren[1,b]

[1]Eindhoven University of Technology, Eindhoven, Netherlands

Abstract

Accurately estimating lettuce growth parameters is crucial for optimizing vertical farming systems, yet existing methods often rely on labor-intensive manual measurements or labeled datasets which can be scarce and costly to acquire. In this work, we propose a novel approach that leverages self-supervised learning techniques to estimate lettuce growth parameters (dry weight, fresh weight, height, diameter, and leaf area) using image data collected throughout the plant's growth cycle. Our methodology consists of a two-part pipeline. First, we implement a self-supervised pre-training step using unlabeled lettuce images obtained at different weeks since seeding. The second part involves fine-tuning the learned weights of a ResNet18 architecture (from the self-supervised pre-training step) on a smaller labeled lettuce dataset for the regression of the 5 growth parameters. We adapt and extend two popular self-supervised learning algorithms, plantSimCLR and plantBT, tailored specifically for plant imagery. Firstly, we propose SimCLR for plants (plantSimCLR) by creating positive and negative pairs based on the time elapsed since seeding. Secondly, we introduce BarlowTwins for plants (plantBT) by applying the redundancy reduction principle to self-supervision. We apply random spatial transformations to the lettuce images to obtain two distorted versions of the original image. The self-supervised pre-training task promotes the representations of the distorted lettuce versions to be close to each other using either contrastive learning (plantSimCLR) or cross-correlation (plantBT). We evaluate the quality of the learned representation against ImageNet pre-trained weights. Our evaluation demonstrates that both plantSimCLR and plantBT provide a more effective starting point for estimating the lettuce growth parameter. Additionally, our method significantly decreases the time and effort needed for manually measuring the lettuce growing characteristics. By leveraging self-supervised learning techniques tailored to plant imagery, our work offers a promising avenue for advancing automated monitoring and optimization of vertical farming systems, ultimately contributing to sustainable and efficient agricultural practices

Keywords: self-supervised, deep learning, lettuce growth, SimCLR, Barlow Twins

[a] E-mail: a.simion.constantinescu@tue.nl ; [b]E-mail: j.vanschoren@tue.nl

INTRODUCTION

Optimal plant growth and product quality depends on the complex interactions among the genetic composition of the plant and environmental factors such as climate, light and substrate. Machine learning has demonstrated remarkable efficacy in accurately modeling these complex relationships. Our innovative challenge lies in comprehensively understanding the intricate dynamics fueling plant growth. Traditionally, understanding such interactions has relied on invasive techniques, such as destructive sampling to measure plant biomass. However, our proposed approach aims to achieve this non-invasively by leveraging image data of the evolving plants.

Convolutional Neural Networks (CNNs) achieve state of-the-art performance on numerous Computer Vision tasks such as image classification (Lee, Won, & Hong, 2020), (Touvron, Vedaldi, Douze, & Jégou, 2020), (Xie, Luong, Hovy, & Le, 2020), video action recognition (Carreira & Zisserman, 2017), (Feichtenhofer, Fan, Malik, & He, 2019), (Stergiou & Poppe, 2020), object detection (Liu, et al., 2016), (Redmon, Divvala, Girshick, & Farhadi, 2016), (Ren, He, Girshick, & Sun, 2015), instance segmentation (Chen, Girshick, He, & Dollár, 2019), (He, Gkioxari, Dollár, & Girshick, 2017), (Kirillov, Girshick, He, & Dollár, 2019) etc. CNNs are the standard backbone architecture for most of Computer Vision domains since the revolution of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012). However, their success is highly dependent on having high-quality labeled data, which is not as largely available for the plant domain. To take advantage of the exiting unlabeled corpora, self-supervised learning (Zisser A., 2018) techniques are explored. The annotation becomes harder as the complexity of crop increases. In this paper we are focusing on lettuce growth parameters estimation without the need of plant scientist data annotation.

Self-supervised pre-training for learning visual representation from images yields very promising results. Methods such as simCLR (Chen, Kornblith, Norouzi, & Hinton, 2020) and Barlow Twins (Zbontar, Jing, Misra, LeCun, & Deny, 2021) are at a competitive level with the fully supervised counterpart for classical computer vision tasks. Compared to the general image domain, plant images poses additional challenges such as different lighting conditions. This paper uses RGB images of lettuces during different growing stages captured on a top-view camera.

Contrastive learning is a type of deep metric learning (Kaya & Bilge, 2019) which aims to learn a function for measuring the similarity between a pair of data points. The function used to distinguish between similar and dissimilar data examples is called a contrastive loss. There are several contrastive loss functions: Siamese loss (Hadsell, Chopra, & LeCun, 2006), Triplet loss (Weinberger, Blitzer, & Saul, 2006), Multi-class N-pair loss (Sohn, 2016), Supervised NT-Xent loss (Khosla, et al., 2020), etc. Self-supervised learning (Zisser A., 2018) is a type of unsupervised learning which uses a pretext task, also called self-supervised task, to generate pseudo-labels from unlabeled data. By optimizing to solve a pretext task which exploits the intrinsic relationship existent in the data, visual representations can be learned. The optimization can be done using a contrastive loss. Our work uses contrastive learning as a way of creating pseudo-labels for the images of lettuce plants during different growing stages.

The self-supervised techniques for image feature learning can be classified based on the type of the pretext task. The generative based methods has the main goal to create realistic images, meaningfully visual representations being learned as a side effect: Denoising auto-encoders (Vincent, Larochelle, Bengio, & Manzagol, 2008), Generative Adversarial Networks (Goodfellow, et al., 2014), Bidirectional GANs (Donahue, Krähenbühl, & Darrell, 2016), Context Encoders (Pathak, Krahenbuhl, Donahue, Darrell, & Efros, 2016), Image colorization (Zhang, Isola, & Efros, 2016). The next class of methods exploits the relationship between patches such as predicting the relative position (Doersch, Gupta, & Efros, 2015), solving Jigsaw puzzles (Noroozi & Favaro, 2016) and learning to count features (Noroozi, Pirsiavash , & Favaro, 2017). Another handcrafted pretext task category is geometric transformation based with Exemplar CNNs (Dosovitskiy, Fischer, Springenberg, Riedmiller, & Brox, 2015) or predicting the image rotation (Gidaris, Singh, & Komodakis, 2018). Finally, contrastive losses are used for the next group of techniques with Contrastive Predictive Coding (Oord, Li, & Vinyals, 2018) being inspired by the Noise Contrastive Estimation used in learning word embeddings (Gutmann & Hyvärinen, 2010). Our work is inspired by two state-of-the-art self-supervised techniques for image feature learning, namely a Simple Framework for Contrastive Learning of Visual Representations (simCLR) (Chen, Kornblith, Norouzi, & Hinton, 2020) and Barlow Twins: Self-Supervised Learning via Redundancy Reduction (Zbontar, Jing, Misra, LeCun, & Deny, 2021). SimCLR uses contrastive learning to maximize agreement between two augmented versions of the same image. The main limitation of this method comes from the size of the batch which determines the number of negative pairs needed for the contrastive loss. Barlow Twins (BT) tackles this problem by applying the redundancy reduction principle to self-supervision, not depending on a contrastive loss that needs a large number of negative pairs.

MATERIALS AND METHODS

General pipeline

As can be seen in *Figure 1*, our methodology consists of a two-part pipeline. First, we implement a self-supervised pre-training step (top part) using unlabeled lettuce images obtained at different weeks since seeding. The second part involves fine-tuning the learned weights of a Convolutional Neural Network (CNN) architecture (from the self-supervised pre-training step) on a smaller labeled lettuce dataset for the regression of the 5 growth parameters. For projecting the lettuce images to a latent space, we use a popular and robust CNN architecture, Resnet18 (He, Zhang, Ren, & Sun, 2016). From the CNN projection, we further propagate the data through a Multi-layer perceptron (MLP) with 2 fully connected (FC) layers and ReLU (Nair & Hinton, 2010) non-linearity in between.

Dataset

The dataset used for the experiments of this paper was generated by Hemming et al. (Hemming, et al., 2021) for the needs of the 3rd International Autonomous Greenhouse Challenge (see here). It consists of RGB images of four varieties of lettuce

(Salanova, Lugano, Satine and Aphylion) grown for seven weeks. The images are 24-bit portable network graphic (png) files with a resolution of 1080x1920. In *Figure 2* a sample of each of the four lettuce cultivars can be seen during a seven weeks growing cycle.

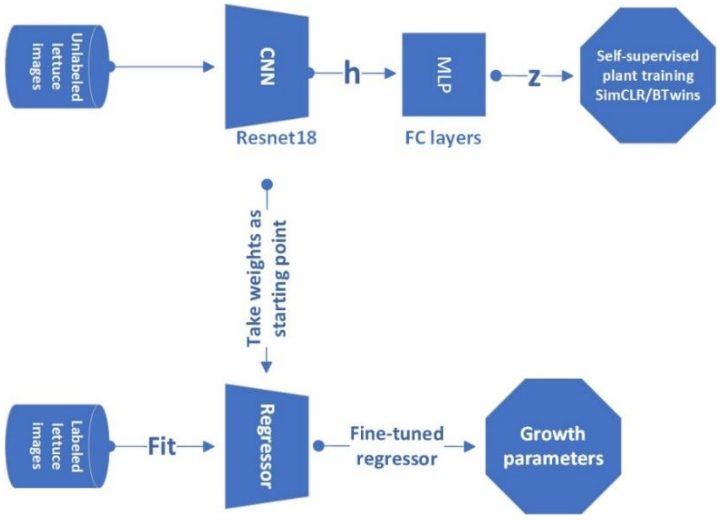

Figure 1 *Overview of our methodology with the top part illustrating the self-supervised pre-training step (unlabeled lettuce images) and the bottom one (small labeled lettuce dataset) showing the fine-tuning regression of the growth parameters*

The most important parameters of lettuce phenotyping that characterize the plant growth are the following : fresh weight (in grams), dry weight measured after a couple of days of drying in a specialized oven (in grams), height obtained from the point where the first leaf was attached to the tallest point of the lettuce (in cm), diameter calculated from the lettuce projection onto a plane (in cm) and leaf area as the area of the surface generated by separating the leaves from the stem (in cm2). All the images from the dataset are labeled according to the 5 growing parameters. An example of these parameters measured by destructive sampling on lettuce from Salanova type harvested during different weeks can be seen in *Table 1*.

Table 1 *Example of the evolution of the growing parameters of a lettuce sample during 7 weeks*

Week	Fresh weight(g)	Dry weight(g)	Diameter (cm)	Leaf area (cm2)	Height (cm)
1	5.2	0.58	17.2	202.7	9.8
2	16.4	1.22	18.5	520.1	6.8
3	69.8	4.16	25.1	1694.3	9.0
4	85.0	5.02	25.0	2008.8	8.0
5	110.0	6.04	28.0	2414.7	12.0
6	133.2	6.84	32.0	3089.2	15.8
7	236.5	11.04	33.5	5348.1	17.0

Figure 2 *RGB images of one sample from each of the four lettuce cultivars (Salanova, Lugano, Satine and Aphylion) taken weekly during a 7-week growing trial. For display purposes, the original images are center cropped. (data link)*

Random spatial data augmentation for plants

We apply a sequence of spatial transformations on the frame level to obtain a different correlated view. The spatial transformation pipeline starts with a random cropping of the original image followed by resizing to the desired picture size, continues with a random horizontal flip, some random color distortions such as color jittering and random grayscaling and finishes with a random Gaussian blur.

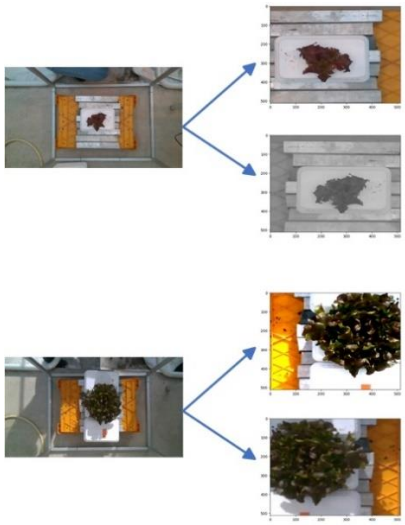

Figure 3 *Example of random spatial augmentations applied to two lettuce images to obtain two distorted versions from each of the two original images*

We use random spatial transformations to the lettuce images to obtain two distorted versions of the original image. An example of applying the spatial transformation pipeline to two lettuce images can be seen in *Figure 3*.

SimCLR for plants (plantSimCLR)

Self-supervised learning is capable of extracting useful information from unlabeled data. Contrastive learning methods are a category of self-supervised algorithms. We first experimented with SimCLR, a contrastive learning approach that learns image representations by maximizing agreement between differently augmented views of the same data example using a contrastive loss. In *Figure 4*, you can see two random spatial augmentations applied on the original image x, t and t'. The contrastive loss is used to maximize the agreement between the positive pairs (zi, zj) given all the negative pairs. The self-supervised contrastive task is used during pre-training to learn useful image representations without having the labels of the images. The self-supervised pre-trained weights are then fine-tuned on a classification/regression task on the available labeled data.

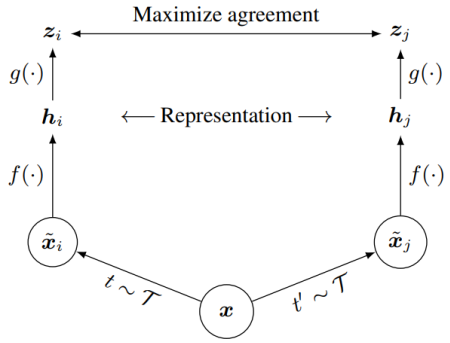

Figure 4 *Overview of "A simple framework for contrastive learning of visual representations" (image courtesy to SimCLR paper)*

The only know measurement without destructive sampling is the time of the plant from the seeding date. Based on this information, positive and negative pairs can be created for plantSimCLR. We define a positive pair as 2 images of lettuces (from different or same varieties) with the same week number since seeding. The negative pairs are the rest of pairs that can be formed within the training batch. The self-supervised pre-training task promotes the representations of the distorted lettuce versions to be close to each other using a contrastive loss.

Barlow Twins for plants (plantBT)

One of the main limitation of the previous discussed self-supervised algorithm (plantsimCLR) is determined by the large batch size needed for the contrastive loss performance. In a batch of N-plants images, 2N augmented images are generated which form N-positive pairs. For each positive pair, there are 2(N-1) negative pairs. The contrastive loss needs a large number of negative pairs to distinguish from the positive ones. The batch size determines the number of negative pairs. The next proposed method tackles this problem.

Another popular self-supervised method for image data is Barlow Twins (BT). It utilizes two identical neural networks, known as twins, to learn representations of data without labels. The key innovation lies in the Barlow Twins loss function, which encourages the learned representations to be invariant to small transformations while preserving useful information. By minimizing the decorrelation between the representations produced by the twins, Barlow Twins can efficiently learn robust and meaningful representations from unlabeled data. This approach has shown promising results in various computer vision tasks, demonstrating its effectiveness in learning useful features without supervision.

We introduce Barlow Twins for plants (plantBT) by applying the redundancy reduction principle to self-supervision. The self-supervised pre-training task promotes the representations of the distorted lettuce versions to be close to each other using cross-correlation. The objective function of plantBT assesses the cross-correlation matrix between embeddings generated by two identical networks when fed with distorted batches of samples, aiming to make this matrix resemble the identity matrix. A schematic of Barlow Twins adapted for plant images is presented in *Figure 5*.

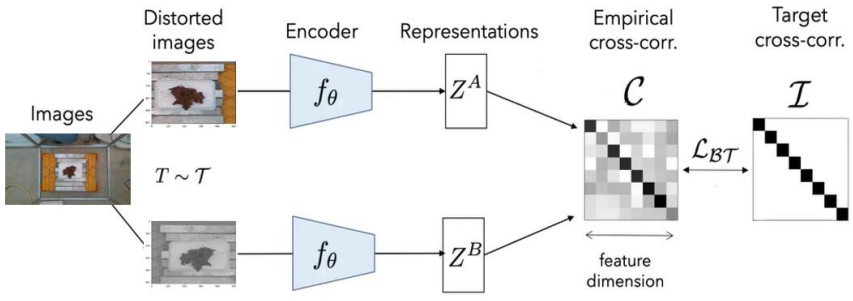

Figure 5 *Overview of plantBT. The distorted versions are obtained using two random spatial augmentations according to the presented transformation pipeline*

RESULTS AND DISCUSSION

Our models are pretrained on the lettuce dataset ignoring the labels. All of the experiments are done using a ResNet18 architecture with only RGB features. Following the linear evaluation protocol, first, the pre-trained weights are used to initialize the ResNet18 regressor architecture. Afterwards, the weights are fine-tuned to train a regressor on the top of the learned representations by using the growth parameters labels of the lettuce dataset. The regressor performance is evaluated in terms of R-Squared (R2) on a smaller held-out testing set. For both self-supervised pre-training and regression, we resize the lettuce images from 1080x1920 to the standard size of 224x224.

ImageNet (Russakovsky, et al., 2014) is a large-scale dataset containing millions of labeled images across thousands of categories. It has been instrumental in advancing computer vision research and deep learning algorithms. Each image is annotated with one of over 20,000 object categories, making it a standard benchmark for computer vision tasks. We evaluate the quality of the learned representation comparing with different starting weights for the ResNet18 regressor model. There are three possibilities: ImageNet supervised pre-trained weights, plantSimCLR self-supervised pre-trained weights and plantBT self-supervised pretrained weights. The results of the comparison in terms of R2 score are displayed in *Table 2*.

Table 2 *R-Squared results for lettuce growth parameters estimation with bolded values being the best score(s) in each category*

Method	Fresh weight(g)	Dry weight(g)	Diameter (cm)	Leaf area (cm2)	Height (cm)
R2 ImageNet	0.93	**0.93**	**0.88**	**0.94**	**0.95**
R2 plantSimCLR	**0.94**	0.92	0.74	0.92	0.68
R2 plantBT	**0.95**	0.91	0.83	0.92	0.71

When comparing with Image Net pre-trained weights, the self-supervised pre-training with plantSimCLR and plantBT generates similar (better only for fresh weight) results for estimating the 5 growth parameters, with the exception of the height parameter.

CONCLUSIONS

In this paper, the challenging task of learning visual features from lettuce images without the need of manually annotated data was tackled. We extended two popular self-supervised algorithms from classical images to plant data, namely plantSimCLR and plantBT. This involved applying a custom spatial transformation pipeline to the lettuce images consisting of a series of random spatial augmentation tailored to plant vision. The evaluation of the learned features was conducted by reporting the R-squared for fine-tuning the ResNet18 regressor on the 5 growth parameters. Our results reveals that both plantSimCLR and plantBT offer a good starting point for estimating lettuce growth parameters when compared to ImageNet pre-trained weights. This could substantially

reduce the time and labor required for manual measurement of lettuce growth traits. Through the application of self-supervised learning methods specifically designed for plant imagery, our research presents a promising direction for enhancing automated monitoring and optimization of vertical farming systems.

The main limitation of our results is given by the data used for our experiments. The lettuce images have only one plant per image. This provided a clear view for the deep learning model, but reduces it applicability to deal with images captured in a vertical farm (where space is limited) that will have multiple lettuce plants with large overlapping between them especially at later growing stages. Further changes needs to be done for both plantSimCLR and plantBT to accommodate this type of images.

ACKNOWLEDGEMENTS

This research is part of the TTW Perspectief programme "Sky High", which is supported by the Dutch Research Council (NWO). See more information here.

Literature Cited

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 6299–6308).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Chen, X., Girshick, R., He, K., & Dollár, P. (2019). Tensormask: A foundation for dense object segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 2061–2069).

Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. *Proceedings of the IEEE international conference on computer vision*, (pp. 1422–1430).

Donahue, J., Krähenbühl, P., & Darrell, T. (2016). Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.
Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. (2015). Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence, 38*, 1734–1747.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. *Proceedings of the IEEE international conference on computer vision*, (pp. 6202–6211).

Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, (pp. 2672–2680).

Gutmann, M., & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, (pp. 297–304).

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, *2*, pp. 1735–1742.
He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, (pp. 2961–2969).

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

Hemming, S. (., de Zwart, H. F., Elings, A. (., bijlaard, m., Marrewijk, B., & Petropoulou, A. (2021). 3rd Autonomous Greenhouse Challenge: Online Challenge Lettuce Images. *3rd Autonomous Greenhouse Challenge: Online Challenge Lettuce Images*. 4TU.ResearchData. doi:10.4121/15023088.v1

Kaya, M., & Bilge, H. Ş. (2019). Deep metric learning: A survey. *Symmetry, 11*, 1066.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., . . . Krishnan, D. (2020). Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Kirillov, A., Girshick, R., He, K., & Dollár, P. (2019). Panoptic feature pyramid networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 6399–6408).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, (pp. 1097–1105).

Lee, J., Won, T., & Hong, K. (2020). Compounding the performance improvements of assembled techniques in a convolutional neural network. *arXiv preprint arXiv:2001.06268*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. *European conference on computer vision*, (pp. 21–37).

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *ICML.*

Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. *European Conference on Computer Vision*, (pp. 69–84).

Noroozi, M., Pirsiavash, H., & Favaro, P. (2017). Representation learning by learning to count. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 5898–5906).

Oord, A. v., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 2536–2544).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, (pp. 91–99).

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, (pp. 1857–1865).

Stergiou, A., & Poppe, R. (2020). Learn to cycle: Time-consistent feature discovery for action recognition. *arXiv preprint arXiv:2006.08247*.

Touvron, H., Vedaldi, A., Douze, M., & Jégou, H. (2020). Fixing the train-test resolution discrepancy: FixEfficientNet. *arXiv preprint arXiv:2003.08237*.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th international conference on Machine learning*, (pp. 1096–1103).

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, (pp. 1473–1480).

Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 10687–10698).

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In M. Meila, & T. Zhang (Ed.), *Proceedings of the 38th International Conference on Machine Learning. 139*, pp. 12310–12320. PMLR. Retrieved from https://proceedings.mlr.press/v139/zbontar21a.html

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. *European conference on computer vision*, (pp. 649–666).

Zisser A., O. G. (2018). Self-Supervised Learning. *Self-Supervised Learning*. Retrieved from https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf