Transfer Learning for Few-Shot Plants Recognition: Antarctic Station Greenhouse Use-Case

Liliya Lemikhova, Sergey Nesteruk, Andrey Somov
Skolkovo Institute of Science and Technology
Moscow, Russia
liliya.lemikhova; sergei.nesteruk; a.somov@skoltech.ru

***Abstract*—In this paper, we apply computer vision for plant recognition at the Antarctic station greenhouse, a training facility for future space colonization missions. Our experiments rely on transfer learning and explore the importance of the pre-training data domain. We show that a common approach of using models pre-trained on the Imagenet dataset can be further improved using publicly available domain-specific datasets. The classification results of 17 plant varieties with the ResNet50 model increase the F-score from 75% to 82% using only 3 training images. We also achieve 78% top-3 accuracy without any training data.**

***Index Terms*—computer vision, few-shot learning, plant phenotyping, transfer learning, zero-shot learning**

I. Introduction

With the fast development of imaging technology [1], computing power [2] and algorithms, computer vision [3] has been widely integrated into the field of agriculture [4] [5]. Such methods are non-invasive, fast and highly precise [6]. Based on the above reasons, researchers are devoted to developing image-based plant phenotyping methods as a supplementary or even alternative to the manual measurement [7] or sensor networks based measurements [8] [9]. Especially it is crucial in places with limited human resources such as Antarctic station (see Fig. 1).

Deep learning methods have shown their great potential and usability for image analysis; however, supervised learning models require large amount of labeled data [10]. Due to connection limitations between the Antarctic station and Europe [11], the amount of available data is limited; so, we have to come up with a solution in a few-shot scenario. The dataset of few hundreds images is not enough to train large modern neural networks with millions of parameters from scratch. Transfer learning is an appropriate technique in such situation [12]. In this case, we train the neural networks on a large dataset and then fine-tune it on the small target dataset. The most popular choice for such dataset is ImageNet dataset [13]. However, for fine-grained agricultural datasets it might be not suitable enough and it has several drawbacks [14]. Most images are object centric and the background is simple in contrast to typical image from agricultural domain [15].

Another option for pretraining is to pretrain on (image, text) pairs, predicting which capture goes with each image. A recent successful implementation of this idea is a contrastive language-image pretraining model introduced in [16]. In addition, this approach performs particularly good for zero-shot prediction when we do not need any data for training.

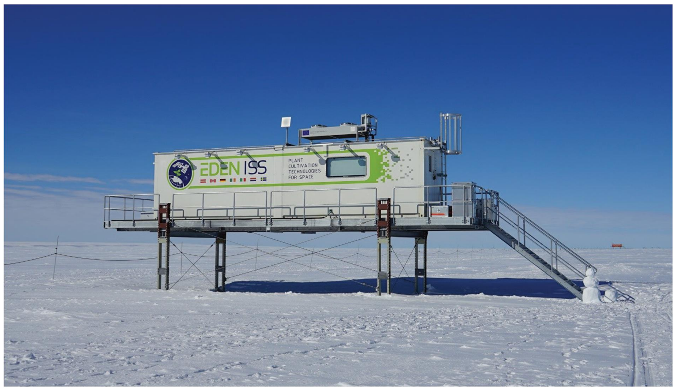

Fig. 1. EDEN ISS Mobile Test Facility.

In this work, we perform a plant recognition task at the Antarctic station greenhouse using computer vision methods. We use transfer learning in our experimental work and investigate the essence of the pre-training data domain. In particular, we demonstrate that the performance achieved by using the models which are pre-trained on the widely used Imagenet dataset can be further increased. It is achieved by relying on the publicly available domain-specific datasets.

The paper is organized as follows: in Section II we introduce the readers to the relevant research works in the area. We discuss the methods used in this research in Section III. In particular, we focus on the datasets and computer vision methods used in this work for image processing. Our experimental results are demonstrated in Section IV where the actual comparative study is carried out. Finally, we provide concluding remarks and discuss our future work in Section V.

II. Related Works

A. Image Analysis of Plant Phenotyping

In [6], the authors extensively reviewed more than 200 papers on plant phenotyping. They highlight data collection, data availability and data analysis as the main challenges of this task.

Many studies use datasets in a controlled environment and with a simple background such as *PlantVillage* dataset [17].

However, such datasets differ greatly from the natural environment [18], and due to differences in light, noise, background the results are much worse when models are tested in real conditions. That is why it is a promising solution for making a robust model to use a dataset with plants taken in various wild places, such as *iNaturalist* dataset [19].

Image annotation is another significant problem for plant phenotyping [20]. On the one hand, it is expensive and time-consuming; on the other hand, it often requires professional knowledge to get accurate labels. The authors conclude that methods based on deep learning fit well to plant phenotyping, but image annotation is the major limiting factor. To overcome this challenge, they suggest developing zero sample and small sample learning. In addition, they emphasize that deep learning methods need a large number of images to fit a large number of parameters; however, deep learning researchers ignore prior domain knowledge, which could reduce the number of learnable parameters, and thus, reduce the sizes of required datasets.

In [21] the authors provided a comprehensive analysis of feature extractors trained with and without supervision on ImageNet and iNaturalist. They found that features produced by standard supervised methods still outperformed those produced by self-supervised approaches. The model pretrained on iNaturalist dataset did not outperform the model pretrained on ImageNet on all tested datasets; but, it showed better performance on the datasets related to nature: Flowers [22], CUB [23] and NABirds [24]. However, the authors did not study the effect of pretraining on iNaturalist for few-shot learning.

B. Few-shot Learning

Few-shot Learning (FSL) is a type of machine learning problem, where the training set contains only a limited number of examples with supervised information for the target task [25]. Usually, few-shot classification means N-way-K-shot classification where the training set contains $I = KN$ examples from N classes each with K examples. When the training dataset does not contain any example with supervised information for the target task, FSL becomes a zero-shot learning problem. Transfer learning [26] transfers knowledge from the source domain/task with a large training dataset, to the target domain/task, where there is a lack of training data. Transfer learning methods are often used in FSL [27] [28] [29].

FSL methods can be categorized into three categories, based on which aspect is enhanced by prior knowledge [25]:

- *Model* to constrain the hypothesis space.
- *Data augmentation* to increase the size of the training set. This includes transforming samples from the training set and similar datasets.
- *Algorithm* to search for best parameterization by providing a good initialization. This includes fine-tuning existing parameters.

In this study, we will focus on the last two groups of methods.

III. Methods

Our goal is to study several scenarios in case of lack of data for training. In this setting, methods can be divided into two groups:

- zero-shot learning – when we do not use samples from the target dataset for training;
- few-shot learning – when we use several samples per class from target dataset.

For few-shot learning, we study different number of samples per class. Thus, we will refer to this setting as to k-shot learning, where k is a non-zero number of training samples from each class.

A. Zero-shot Transfer

There are two approaches that we tried for this scenario.

1) More Data Acquisition : is a common approach when there is no data for training is to find images from similar distribution for each class. If the size of the new dataset is small, we can freeze parameters of the model pretrained on another large dataset and train only last layers. The first disadvantage of this approach is that there is not always publicly available required images. Moreover, it needs additional time and resources to pretrain the model on these new data.

2) CLIP: Contrastive Language-Image Pre-Training (CLIP) is a neural network trained on a variety of (image, text) pairs [16]. It was trained to predict which caption goes with which image on a dataset of 400 million (image, text) pairs collected from the internet. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task.

B. k-shot Learning

As k is usually a small number less than 10, it is impossible to train neural network from scratch. Instead, we use a model with initial weights trained on a large dataset. To tune the model for the target dataset, depending on k, the dataset for pretraining and the target dataset, we can use one of the following scenarios:

- freeze all parameters, add one or several linear layers with activation functions and learn only weights of these final linear layers
- freeze part of deeper layers, where layers identify general features of an image, and fine-tune weights of last layers, which identify more task-specific features of an image.

For the target dataset we conducted experiments with maximum $k = 5$, so we used first strategy for models pretrained on large dataset.

In addition to this method, we further fine-tuned models pretrained for zero-shot learning. For this approach, we used warmup not to degrade learnt weights with too big optimization steps [30]. We also use image augmentation as it has been shown to make models more accurate and stable, especially when there are a small number of training samples [31], [32].

IV. Experimental Results

A. Metrics

To evaluate results, we used weighted F1-score metric [33]. To define F1-score, we first need to define two additional metrics: precision and recall:

$$precision = \frac{TP}{TP + FP}, \tag{1}$$

$$recall = \frac{TP}{TP + FN}, \tag{2}$$

where TP and TN denote the number of positive and negative instances that are correctly classified, while FP and FN denote the number of misclassified negative and positive instances, respectively.

Then F1-score is defined in the following way:

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}. \tag{3}$$

For a multi-class classification problem, we don't calculate an overall F1-score. Instead, we calculate the F1-score per class in a one-vs-rest manner and then find their average weighted by support (the number of true instances for each class).

For zero-shot classification we also calculated top-N accuracy. Top-N accuracy means that the correct class gets to be in the top-N probabilities for it to count as "correct".

B. Datasets

1) ImageNet: is a large visual database with more than 14 million images that have been hand-annotated to classify objects [34]. ImageNet contains more than 20,000 categories.

2) iNaturalist: iNaturalist dataset is a large publicly available fine-grained dataset with more than 800,000 images of more than 5000 species of plants and animals pictures taken in the wild [19].

3) Target dataset: Image collection in the greenhouse is an essential task for guaranteeing the plants growth and health monitoring. The image collection system is the HD color image system, a set of remote-controlled, 4-megapixel, Red-Green-Blue (RGB) Webcams with top-down and side views that monitor the entire greenhouse (see Fig. 2). Some pictures were duplicated and we filtered them. The resulting dataset consists of 228 images of size 1000×1000. There are 16 classes of plants and one extra class for empty slots.

C. Zero-shot Classification

We do not use images from the target dataset for training in this scenario. Instead, we used two methods: getting data from an external open dataset for each class of the target dataset and using the CLIP model in zero-shot mode.

1) Getting more data: for this approach, we needed to find more images close to classes from the target dataset. We tried two options.

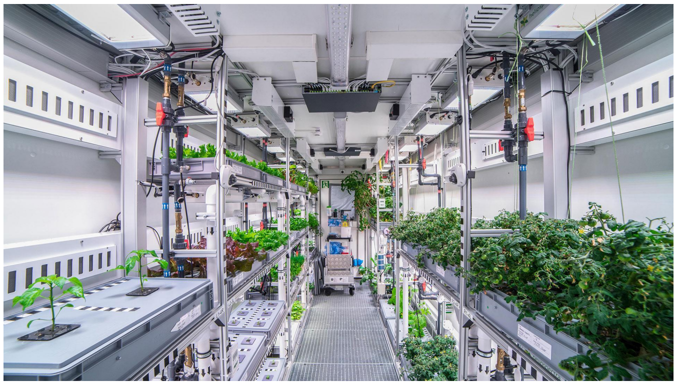

Fig. 2. EDEN ISS Mobile Test Facility inside.

a) Images from iNaturalist dataset: We took 50 images of the same plant family from the iNaturalist dataset for each class of the target dataset except for the class of empty slots. For some classes of the target dataset the plant family from the iNaturalist dataset was the same; thus, to avoid repetition, we assigned images of this family to one of the classes from the target dataset and left others empty.

b) Images retrieved by CLIP: for each class in the target dataset, we formed its textual description in the following form: "{name of class} plants in greenhouse". Then we got 100 images per class whose embeddings are the closest neighbours to the embedding of the corresponding textual description.

After collecting the dataset for pretraining, we need to train the model. First, we froze all layers except the last two convolutional layers of ResNet50 [35] pretrained on ImageNet. Then, we replaced the last fully connected layer with a linear layer with an output dimension equal to the number of classes in the target dataset (17). Finally, fine-tuned parameters to classify collected dataset using cross-entropy loss. We used batch-size of 64, $Adam$ optimizer [36] with initial learning rate $5 \cdot 10^{-4}$, $ReduceLROnPlateau$ scheduler and trained for 20 epochs. We used $RandomResizedCrop(224)$, $RandomRotation$, $ColorJitter$ and $CutOut$ [37] for data augmentation for this approach and for experiments in few-shot classification section.

2) CLIP: for each class, we formed its description in the following form: 'A photo of {name of class} plants in greenhouse'. After that, the model classifies an image depending on the closest class description in embedding space.

The results of these approaches on target dataset are presented in Table I and in Fig. 3.

CLIP with Vision-Transformer [38] as a backbone performs best with all metrics without additional data, which is very convenient. However, the F1-score and top-1 accuracy values are not satisfactory enough to be competitive to few-shot learning results, opposite to what was stated in the paper. This decline of accuracy might be because the pretraining dataset for the CLIP model did not include enough samples from the domain of the target dataset. Thus, future work might be to train CLIP on a dataset that contains samples from the agricultural domain to improve results for zero-shot learning. Additionally, the choice of form for textual description is crucial. The names of plants should be correct, which requires advice from an agriculture specialist. Depending on extra details about context, metrics value alters up to 20%.

Images from iNaturalist of the classes most close to the target dataset show the worst results. Its top-5 accuracy is still lower than top1-accuracy of CLIP with Vision-Transformer as a backbone. The reason is that iNaturalist dataset mostly contains plants in the wild and for example does not contain any kind of lettuce while 6 out of 17 classes from the target dataset are kinds of lettuce. This method also needs lots of time to manually choose appropriate classes.

Images retrieved by CLIP also show relatively poor results. This again may be due to the fact that it retrieves images from the dataset it was trained on. This dataset does not focus primarily on agricultural domain and some of these retrieved images were not appropriate. Some of them differed only with watermarks, some were of classes completely different from required.

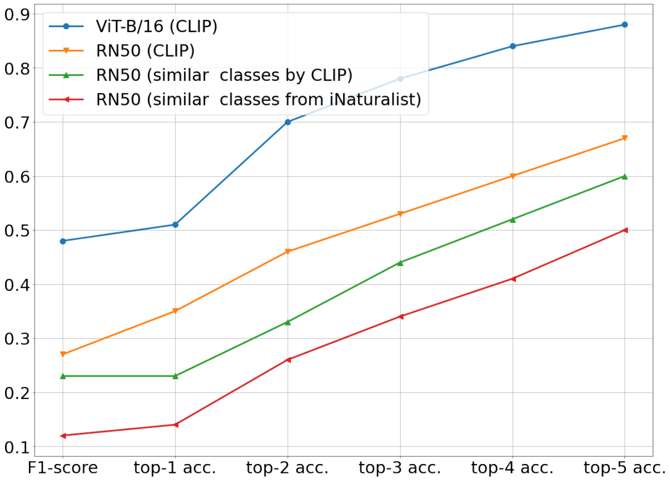

Fig. 3. Metrics for classification in zero-shot scenario.

D. Few-shot Classification

We evaluated results for different numbers k of images per class in the training dataset: 1, 2, 3, 5. For each k, we conducted five experiments with different dataset splits into train and test to estimate the mean and standard deviation of the metric. As a backbone, we used the following pretrained models:

- ResNet50 pretrained on ImageNet as the baseline;
- ResNet50 pretrained on full iNaturalist;
- CLIP with ResNet50 as the backbone;
- CLIP with Vision Transformer-B/16 as the backbone.

For all these models, we froze all parameters and trained only the last fully-connected layer with an output dimension equal to the number of classes. We used batch-size of 64, $Adam$ optimizer with initial learning rate $5 \cdot 10^{-4}$, $ReduceLROnPlateau$ scheduler and trained for 50 epochs.

Also, we further fine-tuned models from experiments for zero-shot learning:

- ResNet50 pretrained on the part of iNaturalist with images of plants of the same family as plants in the target dataset
- ResNet50 pretrained on images of same classes as in target dataset retrieved with CLIP.

We trained these models using the same settings for few-shot learning except for we used warmup for 5 epochs from 0 to $1 \cdot 10^{-4}$ and $StepLR$ scheduler with reducing step every 15 epochs by 5 times.

Results of experiments for few-shot learning are presented in Table II and in Fig. 4.

The iNaturalist dataset consists of images of plants taken in the wild. Moreover, it does not include most of the classes from the target dataset. As a result, it differs strongly from our target dataset. However, pretraining on it shows the best results for all k from 1 to 5 among other pretraining strategies.

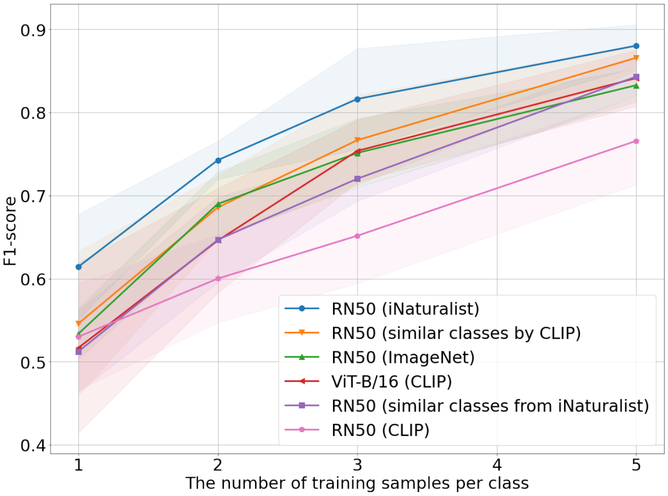

Fig. 4. F1-scores of classification for k-shot scenario depending on number of training examples per class.

The reason why such a successful model as CLIP shows relatively poor results might be that it was trained on dataset that did not contain various enough samples from agricultural domain; thus, it produces feature representations that lack information to distinguish plant varieties. The confusion matrix of CLIP with Vision Transformer as a backbone is presented in the Fig. 5. Label ”5” is most often wrongly classified. The model predicts label ’4’ for five out of eight samples test samples of label ’5’. If we look at images of these two labels, we see that these are two different but still similar kinds of lettuce. Thus, it is better to train on domain-specific datasets than to use a general-case large state-of-the-art model for a fine-grained dataset.

TABLE I
RESULTS FOR ZERO-SHOT CLASSIFICATION

Model	F1-score	top-1 accuracy	top-2 accuracy	top-3 accuracy	top-4 accuracy	top-5 accuracy
RN50 (similar classes from iNaturalist)	0.12	0.14	0.26	0.34	0.41	0.50
RN50 (similar classes by CLIP)	0.23	0.23	0.33	0.44	0.52	0.60
RN50 (CLIP)	0.27	0.35	0.46	0.53	0.60	0.67
ViT-B/16 (CLIP)	**0.48**	**0.51**	**0.70**	**0.78**	**0.84**	**0.88**

TABLE II
F1-SCORES OF k-SHOT CLASSIFICATION WITH $k \in \{1, 3, 5\}$.

Model	k=1	k=3	k=5
RN50 (ImageNet)	0.53 ± 0.03	0.75 ± 0.04	0.83 ± 0.02
RN50 (iNaturalist)	$\mathbf{0.61 \pm 0.06}$	$\mathbf{0.82 \pm 0.06}$	$\mathbf{0.88 \pm 0.03}$
RN50 (similar classes from iNaturalist)	0.55 ± 0.09	0.77 ± 0.05	0.87 ± 0.02
RN50 (similar classes by CLIP)	0.51 ± 0.05	0.72 ± 0.03	0.84 ± 0.02
RN50 (CLIP)	0.53 ± 0.06	0.65 ± 0.06	0.77 ± 0.05
ViT-B/16 (CLIP)	0.52 ± 0.10	0.75 ± 0.04	0.84 ± 0.03

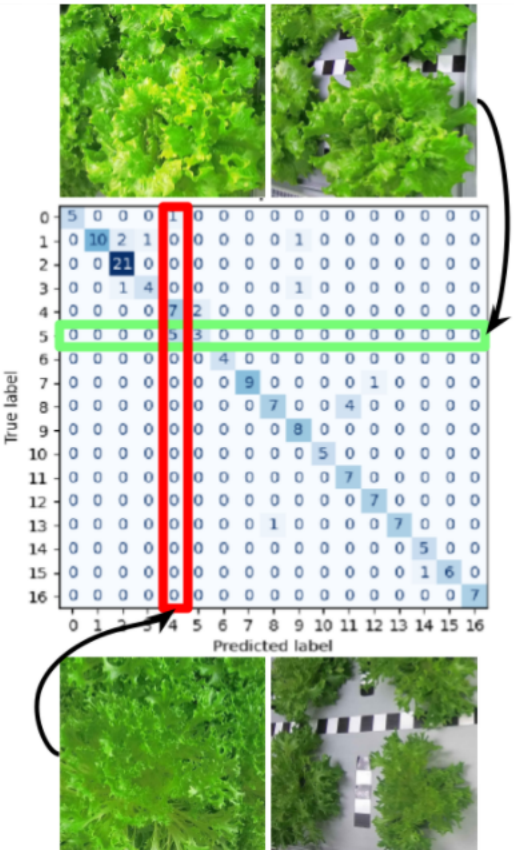

Fig. 5. Confusion matrix of predictions by 5-shot CLIP ViT/16. The model confuses loose-leaf lettuce (label ”4”) and mizuna lettuce (label ”5”).

V. CONCLUSION

We compared different pretraining approaches to solve the plants recognition task for the antarctic station greenhouse in a few-shot scenario, including a zero-shot one. For zero-shot transfer, we used more data acquisition as the first approach and contrastive language-image pretraining as the second one, which showed significantly better performance. Another advantage of the last method is that it is possible to use it without GPUs as the pretrained CLIP is available for download. Moreover, it does not depend on existence of available relevant data and does not require additional time for data processing. For k-shot transfer, we mainly studied relevance of the iNaturalist dataset compared to standard ImageNet for pretraining stage in agricultural domain. In addition, we compared results to state-of-the-art method CLIP. The model pretrained on iNaturalist dataset showed the best results, although a few categories in the target dataset were not among the categories in iNaturalist dataset. Thus, pretraining on a domain closer to agricultural one improves results for few-shot learning. In addition, models pretrained on iNaturalist are available for download; thus, it does not require powerful computation resources to use benefits of this dataset.

Large open datasets such as iNaturalist are a promising alternative to ImageNet for the pretraining stage for fine-grained datasets, especially in precision agriculture. New versions of this dataset are regularly released so that the number of classes and the number of images per class increase. Future work can be to replace ResNet50 with a more sophisticated backbone and to train the model on a more diverse agricultural dataset. Contrastive language-image pretraining for zero-shot learning can be improved by adding more (image, capture) pairs relevant to agricultural domain.

ACKNOWLEDGEMENT

The authors would like to thank German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt e.V.) for sharing data for this research.

REFERENCES

[1] Oleg Sergiyenko, Vera Tyrsa, Wendy Flores-Fuentes, Julio Rodriguez-Quiñonez, and Paolo Mercorelli. Machine vision sensors. *Journal of Sensors*, 2018:1–2, 2018.

[2] Dupont C., Hermenier F., Schulze T., Basmadjian R., Somov A., and Giuliani G. Plug4green: A flexible energy-aware vm manager to fit data centre particularities. *Ad Hoc Networks*, 25(PB):505 – 519, 2015.

[3] Oleg Yu Sergiyenko and Vera V. Tyrsa. 3d optical machine vision sensors with intelligent data management for robotic swarm navigation improvement. *IEEE Sensors Journal*, 21(10):11262–11274, 2021.

[4] Yuzhen Lu and Sierra Young. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, 178:105760, 2020.

[5] Lars Lindner, Oleg Sergiyenko, Moises Rivas-López, Daniel Hernández-Balbuena, Wendy Flores-Fuentes, Julio C Rodríguez-Quiñonez, Fabian N Murrieta-Rico, Mykhailo Ivanov, Vera Tyrsa, and Luis C Básaca-Preciado. Exact laser beam positioning for measurement of vegetation vitality. *Industrial Robot: An International Journal*, 2017.

[6] Zhenbo Li, Ruohao Guo, Meng Li, Yaru Chen, and Guangyao Li. A review of computer vision technologies for plant phenotyping. *Computers and Electronics in Agriculture*, 176:105672, 2020.

[7] Wei Zhao, Xuan Wang, Bozhao Qi, and Troy Runge. Ground-level mapping and navigating for agriculture based on iot and computer vision. *IEEE Access*, 8:221975–221985, 2020.

[8] Roop Pahuja, H.K. Verma, and Moin Uddin. A wireless sensor network for greenhouse climate control. *IEEE Pervasive Computing*, 12(2):49–58, 2013.

[9] Andrey Somov, Ivan Minakov, Alena Simalatsar, Giorgio Fontana, and Roberto Passerone. A methodology for power consumption evaluation of wireless sensor networks. In *2009 IEEE Conference on Emerging Technologies Factory Automation*, pages 1–8, 2009.

[10] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[11] Sergey Nesteruk, Dmitrii Shadrin, Mariia Pukalchik, Andrey Somov, Conrad Zeidler, Paul Zabel, and Daniel Schubert. Image compression and plants classification using machine learning in controlled-environment agriculture: Antarctic station use case. *IEEE Sensors Journal*, 2021.

[12] Zhongjie Yu, Lin Chen, Zhongwei Cheng, and Jiebo Luo. Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12856–12864, 2020.

[13] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From ImageNet to image classification: Contextualizing progress on benchmarks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9625–9635. PMLR, 13–18 Jul 2020.

[15] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[17] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.

[18] Sergey Nesteruk, Dmitrii Shadrin, Vladislav Kovalenko, Antonio Rodríguez-Sánchez, and Andrey Somov. Plant growth prediction through intelligent embedded sensing. In *2020 IEEE 29th International Symposium on Industrial Electronics (ISIE)*, pages 411–416, 2020, 2020.

[19] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] Sergey Nesteruk, Svetlana Illarionova, Timur Akhtyamov, Dmitrii Shadrin, Andrey Somov, Mariia Pukalchik, and Ivan Oseledets. Xtremeaugment: Getting more from your data through combination of image collection and image augmentation. *IEEE Access*, 10:24010–24028, 2022.

[21] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12884–12893, June 2021.

[22] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[23] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[24] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[25] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[26] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. ieee transactions on knowledge and data engineering. *22 (10): 1345*, 1359, 2010.

[27] Samaneh Azadi, Matthew Fisher, Vladimir G Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7564–7573, 2018.

[28] Bo Liu, Xudong Wang, Mandar Dixit, Roland Kwitt, and Nuno Vasconcelos. Feature space transfer for data augmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9090–9098, 2018.

[29] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li Fei-Fei. Label efficient learning of transferable representations across domains and tasks. *arXiv preprint arXiv:1712.00123*, 2017.

[30] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018.

[31] Svetlana Illarionova, Sergey Nesteruk, Dmitrii Shadrin, Vladimir Ignatiev, Mariia Pukalchik, and Ivan Oseledets. Object-based augmentation for building semantic segmentation: Ventura and santa rosa case study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1659–1668, 2021, 2021.

[32] Svetlana Illarionova , Sergey Nesteruk , Dmitrii Shadrin, Vladimir Ignatiev , Maria Pukalchik , and Ivan Oseledets. Mixchannel: Advanced augmentation for multispectral satellite images. *Remote Sensing*, 13(11), 2021.

[33] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[37] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[38] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.