[Reproduce]When Does Self-supervision Improve Few-shot Learning?

Yu-Hsien Su
National Tsing Hua University
112061529
catefanny900812@gmail.com

Shao-Yi Kuo
112061595
dannykuo2000@gmail.com

Hao Wang
112061517
howard.h.wang.23@gmail.com

Jui-Chun Chang
112061594
jimmy3212364@gmail.com

You-Xian Lin
111061804
as111061804@gapp.nthu.edu.tw

1. Introduction

Few-shot learning is a subfield of machine learning that focuses on training models to make accurate predictions or classifications with a small number of labeled data.

Unlike traditional supervised learning, where models typically require large datasets for training, few-shot learning aims to generalize from only a small "few-shot" dataset, which might consist of just a handful of examples or even a single example per class.

1.1. The Importance of Few-shot Learning

Machine learning has made incredible strides in recent years, reshaping the way we approach problems and automating tasks that were once thought impossible. However, one significant challenge has persisted— the need for substantial amounts of labeled training data. In a world where data is often scarce, expensive to collect, or prone to privacy concerns, few-shot learning has emerged as a game-changer.

Consider healthcare, for example, where acquiring a substantial quantity of annotated medical images poses difficulties, and obtaining patients' personal health data raises privacy issues. In such cases, few-shot learning emerges as a promising solution.

To wrap up, few-shot learning stands as a crucial domain in machine learning research and development. Therefore, enhancing the accuracy of few-shot learning will play a pivotal role in effectively tackling real-world issues across a multitude of domains, leading to more dependable and efficient solutions.

1.2. Reproduce and Verify the Few-shot Method

For the aforementioned considerations, we made the decision to reproduce a paper related to few-shot learning. Our choice fell upon "When Does Self-supervision Improve Few-shot Learning?" [9] not only due to the prevalence of self-supervision as a common method in recent years within the few-shot learning domain, but also because it poses pertinent questions frequently encountered in this field.

In the upcoming chapters, we will initially present the methodology employed in this paper, the contributions they claim, and our doubts. Subsequently, we will outline when and how we expect to validate it.

2. Brief Introduction of the Main Paper

In section 2.1, the method used in the target paper will be introduced, and the results will be mentioned in section 2.2.

2.1. The Focus Method of Main Paper

In this paper, they attempt to enhance the performance of few-shot learning by incorporating self-supervised learning (SSL) methods.

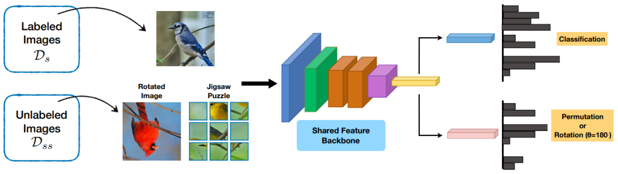


Figure 1. **Combining supervised and self-supervised losses for few-shot learning**

Figure 1 illustrates the framework they employed, integrating SSL into the original meta-learning section.
Denote $\mathcal{D}_s$ as $\{(x_i, y_i)\}_{i=1}^n$ to be an labeled training dataset for meta-learning. And $\mathcal{D}_{ss}$ to be an unlabeled training set for SSL.

The loss function of meta-learning is defined as:

$$\mathcal{L}_s := \sum_{(x_i, y_i) \in \mathcal{D}_s} l(g \circ f(x_i), y_i) + \mathcal{R}(f, g),$$

where f represents feed-forward convolutional network, which maps input x to featured space, g is a classifier, which maps feature to label space, l represents cross-entropy loss function, and R is a $L2$ norm regularization.
The loss function of self-supervised learning is defined as:

$$\mathcal{L}_{ss} := \sum_{x_i \in \mathcal{D}_{ss}} l(g \circ f(\hat{x}_i), \hat{y}_i).$$

For the SSL process, input x would randomly generate $\hat{x}$ and $\hat{y}$ in both method.
The final loss function is defined as:

$$\mathcal{L} := \mathcal{L}_s + \mathcal{L}_{ss}.$$

They adopt two different kinds of SSL methods:

1. **Jigsaw puzzle task loss.** They divided the input image x into 3x3 regions and rearrange them into 35 random combinations, resulting in the rearranged $\hat{x}$ and the target label (index of permutation) $\hat{y}$.
2. **Rotation task loss.** They randomly rotate the input image x by an angle $\theta \in \{0°, 90°, 180°, 270°\}$ to obtain $\hat{x}$, and the target label $\hat{y}$ corresponds to the index of the angle.

2.2. The Focus Results of Main Paper

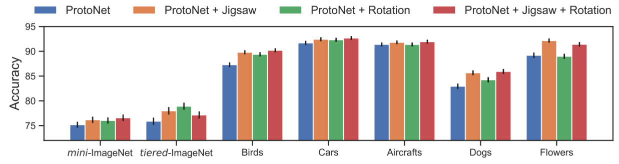


Figure 2. **Benefits of SSL for few-shot learning tasks.**

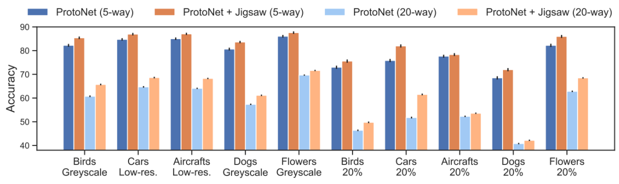
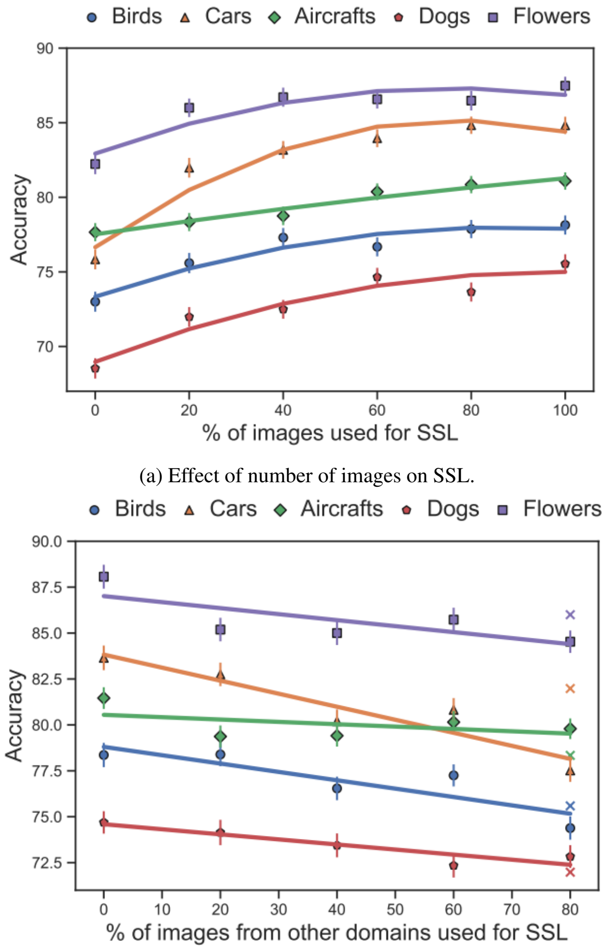


Figure 3. **Benefits of SSL for harder few-shot learning tasks.**



(a) Effect of number of images on SSL.

(b) Effect of domain shift on SSL.

Figure 4. **Effect of size and domain of SSL on 5-way 5-shot classification accuracy.**

In this paper, they claimed that there are benefits of SSL for few-shot learning tasks with no additional training data, as illustrated in figure 2. Furthermore, in figure 3, for those difficult tasks that have been pre-degraded, meaning data that has been pre-processed into grayscale or low-resolution, SSL is more effective.

They also claim that adding more unlabeled data selected by using the proper domain classifier for training SSL can improve performance. In figure 4a, specifically, it presents the relationship between the percentage of images used for SSL in the target domain and the classification accuracy. The result shows that having more unlabeled data from the same domain for SSL improves the performance of the meta-learner. On the contrary,, in figure 4b, when applying domain shift, which means adding images from other domains for SSL, it makes SSL much less effective as the fraction of replacements increases.

In section 4.3 of the paper, they introduced an algorithm for selecting images for self-supervision. They first trained a binary logistic regression domain classifier, using ResNet-

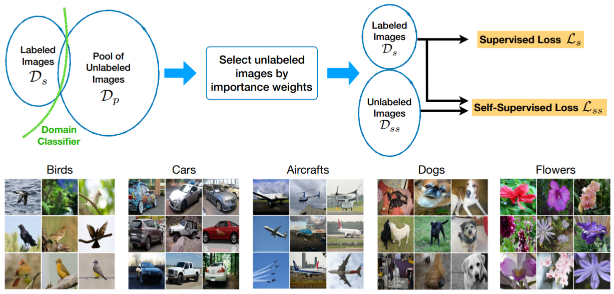

Labeled Images $\mathcal{D}_s$
Pool of Unlabeled Images $\mathcal{D}_p$
Domain Classifier
Select unlabeled images by importance weights
Labeled Images $\mathcal{D}_s$
Unlabeled Images $\mathcal{D}_{ss}$
Supervised Loss $\mathcal{L}_s$
Self-Supervised Loss $\mathcal{L}_{ss}$
Birds
Cars
Aircrafts
Dogs
Flowers

Figure 5. **Overview of domain selection for self-supervision.**

101 [3] as the backbone. Then selected images using this domain classifier for self-supervision learning (figure 5). With random selection, the unlabeled data often has negative effects on the performance, while applying trained parameters on the classifier would improve performance on all datasets.

3. Technical part

3.1. Summary of the technical solution

In the “When Does Self-supervision Improve Few-shot Learning?” [9], several of the claimed contributions by the authors we doubt. Specifically, they claim:

1. **With no additional training data, just adding SSLtask as an auxiliary task can improve the performance.** Since they only test the performance on the common easy open dataset, while the few-shot method is often expected using on the more complex images such as X-ray dataset or human face dataset. We want to evaluate the other real-world datasets.

2. **The SSL method confers substantial advantages when applied to “difficult” datasets.** The paper uses ’low-resolution’ and ’greyscale’ to represent the “difficult” datasets, but we believe there are various other types of “difficult” datasets. Therefore, we want to experiment with different preprocessing datasets to conduct a more comprehensive verification of whether the SSL method yields superior performance on difficult datasets.

3. **The author employs ResNet-18 [3] as the backbone for all experiments and relies on it to support the claims made in the paper.** Given that ResNet-18 is considered an older model by today’s standards, we aim to validate the experiments using more advanced backbones.

4. **Building upon the previous point, this paper resizes all images to 224x224 in order to accommodate the use of ResNet-18 as the backbone.** However, the advanced backbones we intend to employ do not have limitations on image size. Consequently, we aim to investigate whether their asserted contributions remain valid across different image sizes.

3.2. Details of the technical solution

3.2.1 Datasets Preprocessing

As per the referenced study [1], we shall employ datasets derived from diverse domains, including Caltech-UCSD birds [12], Stanford cars [6], FGVC aircraft [7], Stanford dogs [5], and Oxford flowers [8]. Furthermore, we have integrated additional datasets such as CropDiseases [13], EuroSAT [4], ISIC2018 [2], and ChestX [14]. Our approach aligns with the methodology outlined in the reference paper, entailing the partitioning of each dataset into three non-overlapping subsets: base, validation, and novel. The model is trained on the classes in the base set, validated on the classes in the validation set, and tested on the classes in the novel set.

3.2.2 Verification methods

In this chapter, we will elucidate our proposed problem-solving approach for subsection 3.1 and provide an outline of our timeline.

The subsequent solutions will be systematically aligned with the challenges expounded upon in subsection 3.1

1. **Implement on the datasets that better represent real-world conditions.:** We will conduct a validation process to determine if the approach presented in the paper performs well on real-world datasets, as claimed, in order to assess its generalizability.

2. **Expand the definition of “difficult” datasets:** We aim to broaden the definition of challenging datasets in order to more comprehensively investigate the correlation between the method and data complexity.

3. **Translate to the more advanced backbone:** We will conduct validation to ascertain whether the approach presented in the paper continues to yield enhancements when utilized with advanced backbones, such as the transformer [11] or EfficientNet [10].

4. **Resize images in data sets:** We will validate whether the results vary when working with images of varying sizes.

In this study, we employ a distinct and innovative viewpoint compared to the approach outlined in the original paper. This is done to thoroughly assess whether self-supervised learning (SSL) can effectively enhance few-shot learning (FSL).

4. Experiments

4.1. Reproduce original paper setting

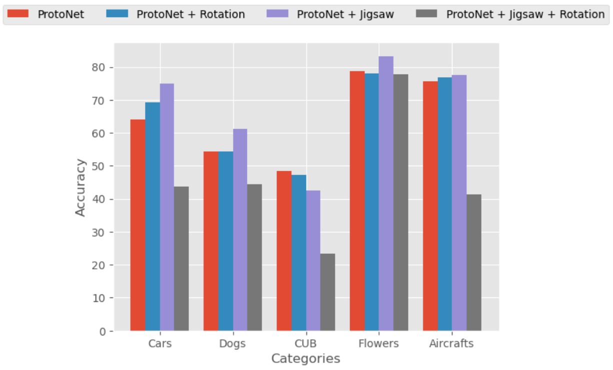


Figure 6. **All reproduce training result in original paper setting.**

The figure referenced as Figure 6 the comprehensive results derived from the methodologies employed in the original research paper. An in-depth discussion and analysis of these findings are presented in the subsequent sections of this document, where we delve into the implications and potential avenues for further exploration based on these outcomes.

4.1.1 Cars dataset

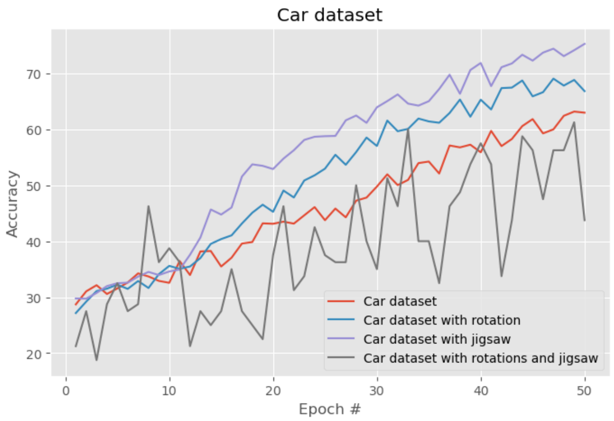


Figure 7. **Training curve of Stanford cars.**

In this section, we utilize the Stanford Cars dataset [6] to evaluate the effectiveness of rotation and jigsaw puzzles in enhancing classification tasks. The training curve results are depicted in Figure 7. The analysis reveals that the accuracy of the baseline model (few-shot learning without self-supervised learning) stands at 64.51%. In contrast, the model incorporating rotation achieves an accuracy of 69.17%, and the model using jigsaw puzzles reaches 74.94%. However, when both self-supervised learning techniques are applied simultaneously, the accuracy drops to 43.68%. This suggests that while the integration of a single self-supervised learning method can improve performance, combining both adversely impacts the model's effectiveness on the Stanford Cars dataset.

4.1.2 Dogs dataset

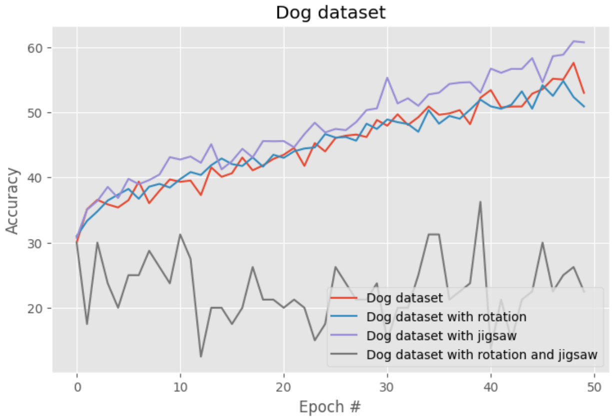


Figure 8. **Training curve of Stanford dogs .**

We use the Stanford Dogs dataset to evaluate the effectiveness of rotation and jigsaw puzzles in enhancing classification tasks.The training results are illustrated in Figure8. The result indicates that the accuracy of the baseline model is 54.33%.When the model incorporates rotation, the accuracy is 54.34%, showing minimal improvement.However,the model with jigsaw puzzles, achieves an accuracy of 61.22%. These results indicate that jigsaw puzzles can indeed improve the model's accuracy when using the Stanford Dogs dataset, while rotation, on the other hand, does not.Furthermore, when we simultaneously apply both SSL techniques on the model, the accuracy of the model drops to 27.49%.This suggests that using both SSL techniques simultaneously on the model results in a negative impact on its performance.

4.1.3 Birds dataset

Similar to our approach with the previous dataset, we applied the same method to the Caltech-UCSD Birds-200-2011 dataset. The outcomes of this training are shown in Figure 9. Our findings reveal that the baseline model attained an accuracy of 48.46%. Interestingly, the introduction of any SSL method led to a decrease in performance. This result diverges from the assertions made in the original paper. A potential explanation for this discrepancy could

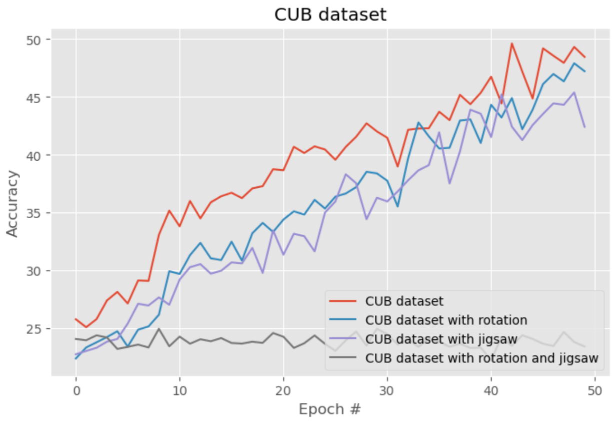


Figure 9. **Training curve of CUB-200-2011.**

be that CUB dataset become are more complecated to others. It is possible that without sufficient training epochs, the advantages of incorporating SSL are not readily apparent.

4.1.4 Aircrafts dataset

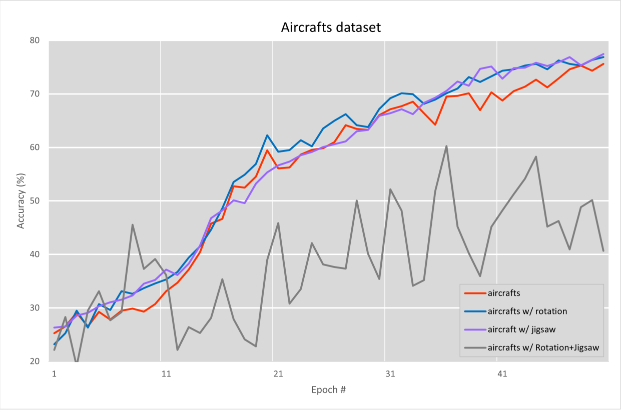


Figure 10. **Training curve of fgcv-aircrafts.**

We also utilized the fgcv-aircrafts dataset as one of the evaluation criteria for assessing the effectiveness of rotation and jigsaw puzzles in enhancing classification tasks. The training results are shown in Figure 10, where the baseline performance is observed to be 75.62%. Upon incorporating rotation, the performance slightly improves to 76.9%, and with the jigsaw puzzles, it further increases to 77.43%. However, when both effects are introduced simultaneously, the performance drops to 41.32%. Therefore, combining both rotation and jigsaw puzzles may not be suitable for this dataset.

4.1.5 Flowers dataset

In Oxford flowers [8] dataset, the accuracy of the baseline model stands 78.73%, which is shown in figure11, and the model with rotation only arrives 78.01%, the performance is 0.72% lower compared to the baseline, mainly due to insufficient training epochs and the SSL rotation not fully leveraging its potential. However, utilizing the SSL jigsaw model has shown significant improvement, with an accuracy boost of 4.53%. It appears that self-supervised learning, when effectively utilized, can yield excellent results on the flowers dataset.

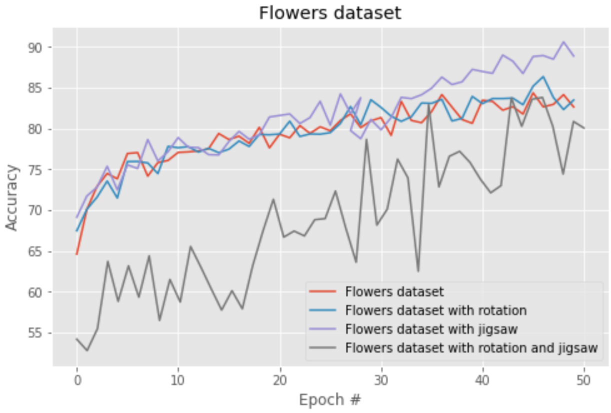


Figure 11. **Training curve of Oxford flowers.**

4.2. Verify ideas by novel setting

4.2.1 Different dataset

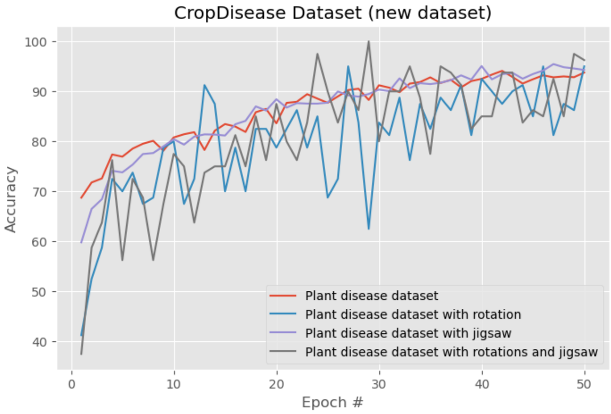


Figure 12. **Training curve of Crop Disease.**

We apply the CropDiseases dataset [13] to assess the impact of rotation and jigsaw puzzles on enhancing classification tasks. The results of the training curve are illustrated in Figure 12, and validation accuracy is also shown in Figure

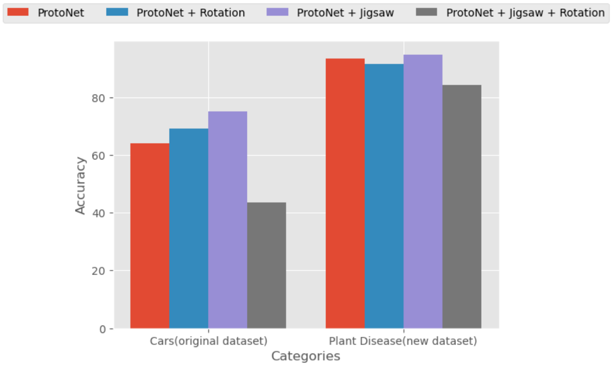


Figure 13. **New dataset and original dataset accuracy**

13. Our analysis indicates that the baseline model, which employs few-shot learning without self-supervised learning techniques, achieves an accuracy of 93.33%. When rotation is added to the model, the accuracy slightly decreases to 91.55% in validation set, even thought training curve result is better. Conversely, the integration of jigsaw puzzles results in a higher accuracy of 94.81%. However, combining both self-supervised learning methods results in a reduced accuracy of 88.85%. Base on validation result and training curve, it demonstrates that their concurrent application does not yield improvements in few-shot learning tasks on the CropDiseases dataset.

4.2.2 Smaller image

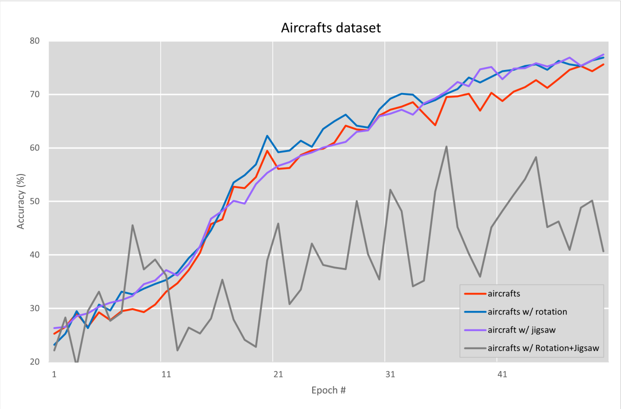


Figure 14. **Training curve of CUB-200-2011 smaller image size.**

4.2.3 Bigger image

We use the AirCrafts dataset to evaluate the impact of increasing image size to 448*448 on enhancing classification tasks. The results of the training curve are shown in Figure 15 , and the validation accuracy is also presented in Figure 16 . Our analysis indicates that when we enlarge the images, the baseline model's accuracy drops to only 30.41% When rotation is added to the model, accuracy slightly decreases to 30.13%, while jigsaw puzzles lead to higher accuracy, reaching 30.51% and when both effects are introduced simultaneously, the performance drops to 29.54%. Compared to the original size, this suggests that a larger image size does not yield better results.

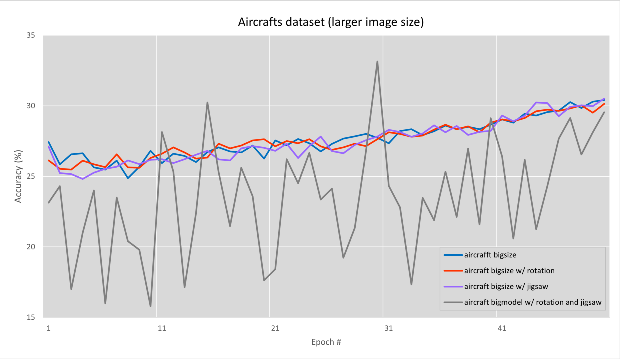


Figure 15. **Training curve of fgcv-aircrafts larger image size.**

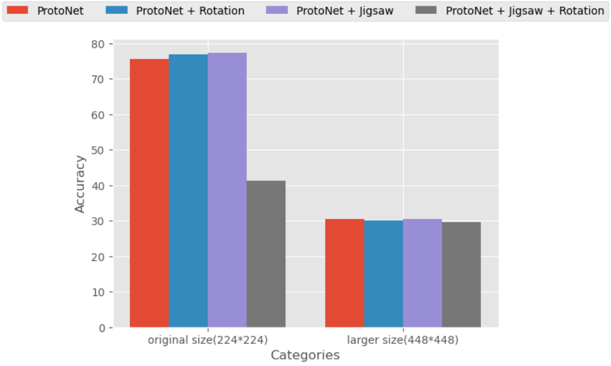


Figure 16. **larger image size and original image size accuracy**

4.2.4 Gray image

We introduce grayscale to the Stanford Dogs dataset to assess the impact of rotation and jigsaw puzzles on enhancing classification tasks.The training results are illustrated in Figure 17,and validation accuracy is also shown in Figure 18.The Figure 17 indicates that the accuracy of the baseline model is 54.34%.Unlike the model without grayscale, the model incorporating rotation or jigsaw puzzles improves the accuracy ,and the accuracy of both is 57.70% and 60.10% respectively.Similarly,using rotation and jigsaw puzzles on the model results in a decrease in accuracy,and its accuracy is 29.20%.From Figure 18, we can conclude that introducing grayscale to the Stanford Cars dataset does not have a negative impact on the effectiveness of rotation and jigsaw

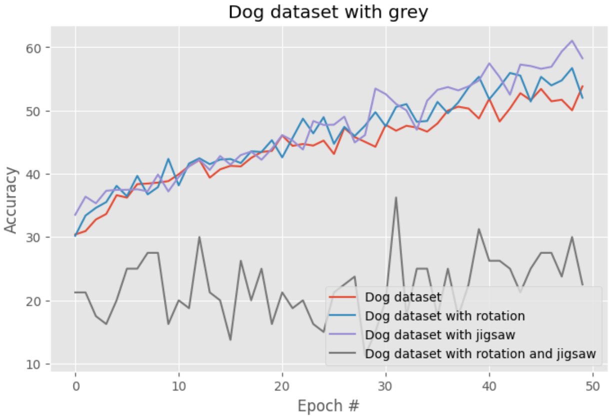


Figure 17. **Training curve of Stanford dogs with gray.**

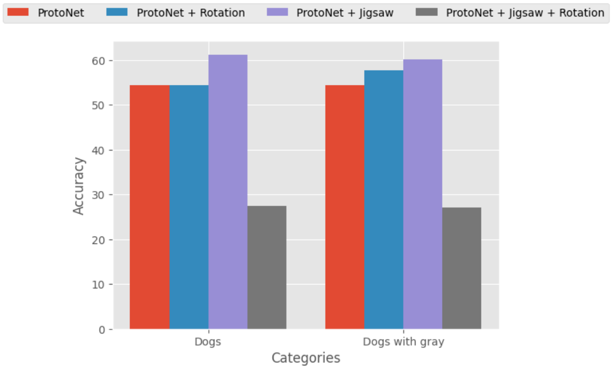


Figure 18. **Dog and dog with gray accuracy.**

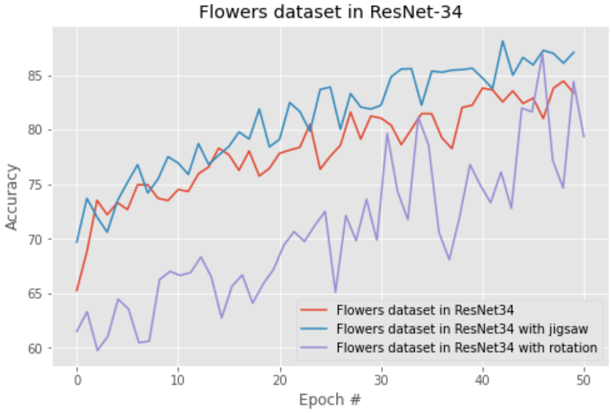


Figure 19. **Training curve of Oxford flowers in ResNet-34.**

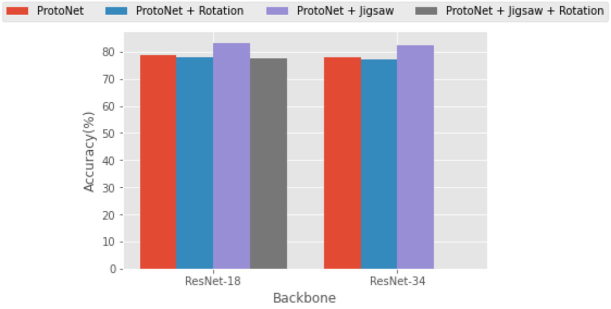


Figure 20. **Flowers and flowers in ResNet-34 accuracy.**

puzzles in enhancing classification tasks.

4.2.5 Different backbone model

We also experimented with different backbone modules, and in this case, we adopted ResNet-34. In figure 19, ResNet-34 achieved an accuracy of 77.73% in the baseline, slightly below the 78.73% baseline of ResNet-18. This may be attributed to the notion discussed in class: larger models require more extensive data. Additionally, the fact that we did not utilize pre-trained weights becomes more evident given the relatively insufficient size of the flowers dataset, consisting of only 8189 files. In terms of SSL performance, the jigsaw method yielded a 4.5% improvement in accuracy. The magnitude is similar to that observed with ResNet-18, demonstrating the adaptability of the SSL method employed in this study across different modules. And we also compare the result in figure 20.

5. Conclusion

In this final project, we attempted to reproduce a paper that utilized self-supervised learning to enhance few-shot learning. Apart from validating the reported effects within the paper's content, we also designed numerous methods to test its generality. These methods included testing on new datasets, substituting with more complex backbones, and using different sizes of images as inputs, among others. The test results partially aligned with the arguments presented in the paper, while some discrepancies were observed. In the original paper, the addition of self-supervised learning consistently led to improved performance. However, in our practical execution, challenges such as insufficient training epochs, difficulty in training the self-supervised learning component, and code-related issues were encountered. Instances where the outcomes did not quite match the arguments in the paper were typically associated with these challenges. Nevertheless, on a broader scale, we can still affirm that incorporating additional self-supervised learning is beneficial for model performance and is applicable across different datasets and backbones.

References

[1] Arjun Ashok and Haswanth Aekula. When does self-supervision improve few-shot learning?-a reproducibility report. In *ML Reproducibility Challenge 2021 (Fall Edition)*, 2022. 3

[2] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 3

[3] 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. 3

[4] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian 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*, 12(7):2217–2226, 2019. 3

[5] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011. 3

[6] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 3, 4

[7] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3

[8] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006. 3, 5

[9] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *European conference on computer vision*, pages 645–666. Springer, 2020. 1, 3

[10] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[12] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 3

[13] Dale R Walters, Jaan Ratsep, and Neil D Havis. Controlling crop diseases using induced resistance: challenges for the future. *Journal of experimental botany*, 64(5):1263–1280, 2013. 3, 5

[14] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 3