# Evaluating Pretraining Methods for Deep Learning on Geophysical Imaging Datasets

**James Chen**
Kirby School
jyc575@gmail.com

## Abstract

Machine learning has the potential to automate the analysis of vast amounts of raw geophysical data, allowing scientists to monitor changes in key aspects of our climate such as cloud cover in real-time and at fine spatiotemporal scales. However, the lack of large labeled training datasets poses a significant barrier for effectively applying machine learning to these applications. Transfer learning, which involves first pretraining a neural network on an auxiliary "source" dataset and then finetuning on the "target" dataset, has been shown to improve accuracy for machine learning models trained on small datasets. Across prior work on machine learning for geophysical imaging, different choices are made about what data to pretrain on, and the impact of these choices on model performance is unclear. To address this, we systematically explore various settings of transfer learning for cloud classification, cloud segmentation, and aurora classification. We pretrain on different source datasets, including the large ImageNet dataset as well as smaller geophysical datasets that are more similar to the target datasets. We also experiment with multiple transfer learning steps where we pretrain on more than one source dataset. Despite the smaller source datasets' similarity to the target datasets, we find that pretraining on the large, general-purpose ImageNet dataset yields significantly better results across all of our experiments. Transfer learning is especially effective for smaller target datasets, and in these cases, using multiple source datasets can give a marginal added benefit.

## 1 Introduction

As raw geophysical data is collected in ever-increasing volumes, there is a need for automated tools to extract useful information to better understand and monitor climate change. Machine learning has the potential to help us analyze climate data accurately, quickly, and at finer spatiotemporal scales than standard methods [2, 3]. For example, there has been significant recent interest in using computer vision models to classify cloud types from images [11, 27, 28]. The presence of different types of clouds has important implications for climate change because clouds have diverse impacts on radiative forcing: certain cloud structures enhance warming by trapping heat, while others mitigate warming by reflecting heat away [4]. With complex feedbacks between cloud characteristics and warming surface temperatures, clouds play a vital role in the sensitivity of the climate to changes in $CO_2$ concentration [6, 16, 26]. At the same time, the precise impact of clouds on climate change is difficult to model [27]; the sixth IPCC assessment report on climate change states "clouds remain the largest contribution to overall uncertainty in climate feedbacks" [21]. Therefore, there is great utility in applying machine learning to automatically classify cloud images, allowing scientists to analyze clouds at finer spatiotemporal scales and continuously monitor changes in cloud cover and its impacts on the climate.

The success of machine learning depends on large *labeled* datasets such as the ImageNet dataset which contains over a million images scraped from the internet [8]. While raw geophysical images

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

are plentiful from ground-based and remote sensors, there is a relative dearth of labeled images: many academic datasets contain only hundreds or thousands of labeled images [3, 9, 18, 28, 30]. In situations where labeled data is limited, transfer learning can provide a simple and effective method of training accurate machine learning models [15, 17]. This method involves first pretraining a model on an auxiliary dataset known as the "source dataset." Then the model is further trained on the "target dataset" of interest, a process known as finetuning. This allows us to transfer the patterns learned in the source dataset to augment the training of the target dataset. Transfer learning has been successfully used to train accurate machine learning models from limited labeled data for applications ranging from cloud classification to weather forecasting to land use classification to prediction of El Niño-Southern Oscillation events [3, 13, 20, 24]. However, the use of transfer learning varies widely across machine learning models for geophysical applications. Some models are not pretrained on a source dataset [11, 18], some are pretrained on the standard machine learning ImageNet dataset [7, 20, 30], and others are pretrained on task-specific source datasets such as other geophysical imaging datasets or simulated data [13, 24, 28].

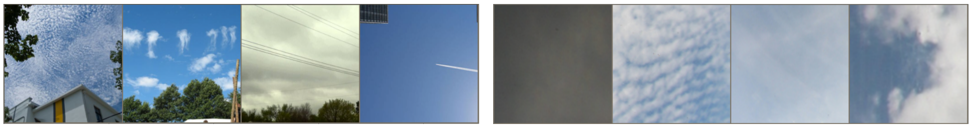
We systematically evaluate how the choice of source dataset impacts transfer learning across three geophysical tasks: cloud classification, cloud segmentation, and aurora classification. We compare the accuracies of models with no pretraining, models pretrained on the general-purpose ImageNet dataset, and models pretrained on task-specific geophysical datasets (e.g. other cloud classification datasets). Results are varied across prior work on evaluating pretraining in other domains from medical imaging to law texts: while pretraining on general-purpose datasets is common, sometimes similar results can be obtained with no pretraining or improvements can be gained by pretraining on domain-specific datasets [1, 19, 23, 29].

In the context of geophysical imaging datasets, we find that transfer learning can significantly improve the performance of the models, up to an increase in test accuracy of 10 percentage points. Across the board, pretraining on the ImageNet dataset provides more added benefit than pretraining on smaller but more related task-specific source datasets. This finding indicates that the utility of transfer learning is to some extent task-agnostic; the benefit of pretraining on a dataset the size of ImageNet outweighs the fact that the images are of everyday objects rather than of clouds or auroras. We further experiment with multiple steps of transfer learning where we pretrain on multiple source datasets. In general this yields little to no additional benefit over simply pretraining on a single source dataset, but there are a few instances where this method provides small increases in accuracy. In all cases, the benefits of transfer learning are most apparent for smaller target datasets. We hope this work will give scientists insight into the transfer learning pipelines that will get the most out of small geophysical imaging datasets and thus aid in automating analysis and monitoring of our climate.

## 2 Methods

We focus on three geophysical imaging tasks: cloud classification, cloud segmentation, and aurora classification. Machine learning models for cloud classification and segmentation in particular have great potential for improving our understanding of the climate and of climate change. As general image classification and image segmentation have been intensely studied in computer vision, we make use of existing machine learning models and large benchmark datasets. For each task, we use several "target" datasets for which we want to develop accurate machine learning models. To better understand the utility of transfer learning for geophysical imaging tasks, we evaluate the accuracy of machine learning models pretrained on a variety of "source" datasets. The source datasets include the other target datasets for the same task (e.g. we pretrain on one cloud classification dataset and finetune on another cloud classification dataset) as well as image processing datasets from other fields. In addition, for each task we experiment with multiple transfer learning steps in which we pretrain on multiple source datasets in sequence (e.g. pretrain on ImageNet, then pretrain on a source cloud dataset, then finetune the model on the target cloud dataset). Further details on the tasks and datasets are given later in this section and in Table 2 (Appendix A).

We implement all of our transfer learning experiments in Python using the PyTorch framework [22]. Each experiment is averaged over 10 trials, and we report one standard deviation. For all of our classification tasks, we use the ResNet-18 model architecture with a final softmax layer [14]. For all of our segmentation tasks, we use the U-Net architecture with a final softmax layer [25]. The models are trained using stochastic gradient descent with momentum. No layers are frozen: all weights are updated in the finetuning stage.



(a) CCSN (cirrocumulus, altocumulus, nimbostratus, contrails)

(b) SWIMCAT (thick dark clouds, patterned clouds, veil clouds, thick white clouds)

Figure 1: Sample Cloud Classification Images

**Target Datasets** We test two target datasets for the task of cloud classification. CCSN [28] contains 2,543 images of 11 classes of clouds: cirrus, cirrostratus, cirrocumulus, altocumulus, altostratus, cumulus, cumulonimbus, nimbostratus, stratocumulus, stratus, and contrails. SWIMCAT [9] contains 784 images collected by an all-sky camera in Singapore of 5 classes of clouds: clear sky, patterned clouds, thick dark clouds, thick white clouds, and veil clouds.

We test two target datasets for the task of cloud segmentation. SWIMSEG [10] contains 1,013 images of clouds with corresponding binary segmentation maps indicating which pixels represent clouds and which do not. SWINSEG [12] contains 115 nighttime images of clouds with corresponding binary segmentation maps. Accuracy for cloud segmentation refers to pixel-wise accuracy.

We test two target datasets for the task of aurora classification. The Kiruna dataset [18] contains 3,846 images collected by an all-sky camera near Kiruna, Sweden of 7 classes of auroras: breakup, colored, arcs, discrete, patchy, edge, and faint. The Yellow River 2 (YR2) dataset [30] contains 8,001 images collected by an all-sky camera at the Yellow River Station of 4 classes of auroras: arc, radiation corona, hot spot corona, and drapery corona.

**Source Datasets** For each target cloud classification dataset, we pretrain on the other cloud classification target dataset, along with ImageNet, a commonly used image classification dataset with over a million images of 1,000 classes ranging from animals to everyday objects. For example, when testing CCSN as the target dataset, we evaluate the performance of no transfer learning, transferring from ImageNet, and transferring from SWIMCAT. For each target cloud segmentation dataset, we try pretraining on the other cloud segmentation target dataset, as well as a third cloud segmentation dataset called SWINySEG [11] which contains 6,768 daytime and nighttime images of clouds with corresponding binary segmentation maps. The U-Net image segmentation model was originally designed for medical imaging applications, so we also experiment with pretraining on the LGG dataset, a brain MRI segmentation dataset with 7,858 images [5]. For each target aurora classification dataset, we pretrain on the other aurora classification target dataset, along with ImageNet. We also pretrain on the Yellow River 1 (YR1) dataset [30] which contains 1,200 images with the same classes as YR2. We do not use YR1 as a target dataset because even with no pretraining, ResNet-18 achieves 100% test accuracy, so there is no room for improvement.

# 3 Results

| Source Dataset | Target Dataset | Train Accuracy (%) | Test Accuracy (%) |
|---|---|---|---|
| None | CCSN | 51.00(±1.88) | 32.87(±1.28) |
| ImageNet | CCSN | 92.59(±0.36) | **40.33(±1.48)** |
| SWIMCAT | CCSN | 67.50(±2.82) | 33.71(±0.96) |
| ImageNet → SWIMCAT | CCSN | 93.21(±0.44) | 38.34(±2.23) |
| None | SWIMCAT | 96.17(±0.89) | 87.12(±2.12) |
| ImageNet | SWIMCAT | 98.18(±0.65) | 95.32(±1.67) |
| CCSN | SWIMCAT | 96.74(±0.92) | 88.33(±3.77) |
| ImageNet → CCSN | SWIMCAT | 98.64(±0.28) | **97.63(±0.95)** |

Table 1: Cloud Classification Results (best result for each target dataset is bolded and one standard deviation is displayed in parentheses).

**Varying Source Datasets** Table 1 shows very pronounced differences in performance between different source datasets for cloud classification. The CCSN dataset has significant background information (see Figure 1), making it difficult for machine learning models to identify relevant features and resulting in lower test accuracy with the model potentially overfitting to the background. For both CCSN and SWIMCAT as target datasets, pretraining on ImageNet significantly outperforms

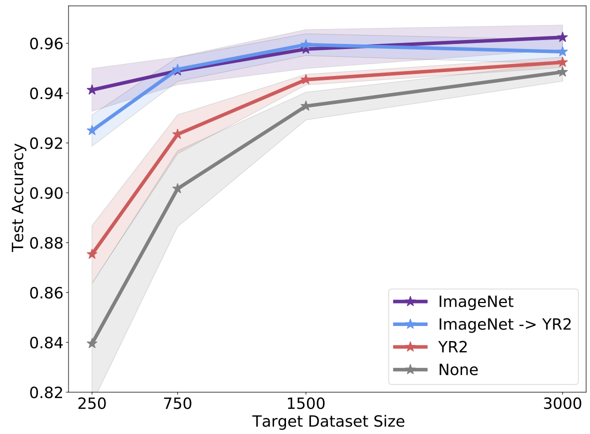



Figure 2: Comparison of source datasets for Kiruna with varying target dataset size. Shading indicates one standard deviation.

pretraining on the other cloud classification dataset and increases accuracy by more than 7% over no transfer learning in both cases. Interestingly, pretraining on ImageNet and CCSN and then finetuning on SWIMCAT adds an additional 2% improvement over just pretraining on ImageNet. While pretraining on multiple source datasets does not always yield an added benefit, it can modestly improve accuracy for small target datasets such as SWIMCAT.

As shown in Table 3 (in Appendix B), even though the larger SWINySEG dataset achieves a relatively low accuracy of 59.8%, it is still the most effective source dataset for cloud segmentation (ImageNet is not applicable in this case as it is a classification dataset). For both SWIMSEG and SWINSEG as target datasets, pretraining on SWINySEG increases accuracy by approximately 1.5%. Larger source datasets are once again the most effective, even if the model did not originally achieve high test accuracy on the source dataset.

For aurora classification (see Table 4 in Appendix C), ImageNet consistently outperforms smaller, domain-specific source datasets. However, with aurora datasets, we achieve relatively good performance even without transfer learning ($> 90\%$) and the target datasets are relatively large compared to those for cloud classification and segmentation, so the advantage of transfer learning is not as pronounced. For target dataset Kiruna, pretraining on ImageNet increases accuracy by 1.4% over no pretraining. For target dataset YR2, the accuracy increases by 3%. For both target datasets, multiple transfer learning steps lead to worse performance than simply pretraining on ImageNet.

**Varying Target Dataset Size** In Figure 2, we show the results of varying the size of target dataset by randomly subsampling training sets of sizes 3000, 1500, 750, and 250 from the Kiruna dataset. For each of these training set sizes, we compare no transfer learning to pretraining on three different source datasets: ImageNet, YR2, and ImageNet $\rightarrow$ YR2. There is a consistent order in performance across the source datasets with ImageNet and ImageNet $\rightarrow$ YR2 performing similarly, followed by YR2, and then None. The differences become much more pronounced as the training set shrinks. For example, with a training set size of 3000, there is only a 1% difference in test accuracy between the best and worst performing source dataset (ImageNet and YR2, respectively). However, when using a training set of size 250, this difference in test accuracy grows to 6.6%, with ImageNet achieving 94.1% accuracy and YR2 achieving only 87.5% accuracy. This supports the idea that choosing the right source dataset is especially important for small target datasets.

## 4 Conclusion

With huge volumes of geophysical imaging data and a relative dearth of labeled images, transfer learning is a useful tool to effectively apply deep learning to analysis and monitoring of our climate at a more granular scale. Choosing the right (often the largest) source dataset for pretraining has a significant impact on the utility of transfer learning, especially for smaller target datasets. Next steps might include (1) evaluating if transfer learning behaves similarly for modalities other than all-sky images such as remote sensing or non-image data; and (2) exploring how to incorporate transfer learning into hybrid pipelines that combine deep learning with physics-based models.

## Acknowledgements

The author thanks Dr. Kara Lamb (Columbia University), Harikrishna Kuttivelil (UC Santa Cruz), and Erik Perkins (Kirby School) for their mentorship. The author also thanks Climate Change AI and the UC Santa Cruz Science Internship Program.

## References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, November 2019. Association for Computational Linguistics.

[2] Karianne J. Bergen, Paul A. Johnson, Maarten V. de Hoop, and Gregory C. Beroza. Machine learning for data-driven discovery in solid Earth geoscience. *Science*, 363(6433), 2019.

[3] Tom Beucler, Imme Ebert-Uphoff, Stephan Rasp, Michael Pritchard, and Pierre Gentine. Machine Learning for Clouds and Climate (Invited Chapter for the AGU Geophysical Monograph Series "Clouds and Climate"). 05 2021.

[4] Sandrine Bony, Bjorn Stevens, Dargan M. W. Frierson, Christian Jakob, Masa Kageyama, Robert Pincus, Theodore G. Shepherd, Steven C. Sherwood, A. Pier Siebesma, Adam H. Sobel, Masahiro Watanabe, and Mark J. Webb. Clouds, circulation and climate sensitivity. *Nature Geoscience*, 8(4):261–268, March 2015.

[5] Mateusz Buda, Ashirbani Saha, and Maciej A. Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*, 109:218–225, 2019.

[6] Paulo Ceppi and Peer Nowack. Observational evidence that cloud feedback amplifies global warming. *Proceedings of the National Academy of Sciences*, 118(30), 2021.

[7] Lasse B. N. Clausen and Hannes Nickisch. Automatic Classification of Auroral Images From the Oslo Auroral THEMIS (OATH) Data Set Using Machine Learning. *Journal of Geophysical Research: Space Physics*, 123(7):5640–5647, 2018.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[9] Soumyabrata Dev, Yee Hui Lee, and Stefan Winkler. Categorization of cloud image patches using an improved texton-based approach. In *in Proc. IEEE International Conference on Image Processing (ICIP)*, 2015.

[10] Soumyabrata Dev, Yee Hui Lee, and Stefan Winkler. Color-Based Segmentation of Sky/Cloud Images From Ground-Based Cameras. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(1):231–242, January 2017.

[11] Soumyabrata Dev, Atul Nautiyal, Yee Hui Lee, and Stefan Winkler. CloudSegNet: A Deep Network for Nychthemeron Cloud Image Segmentation. *IEEE Geoscience and Remote Sensing Letters*, PP:1–5, 05 2019.

[12] Soumyabrata Dev, Florian M. Savoy, Yee Hui Lee, and Stefan Winkler. Nighttime sky/cloud image segmentation. *arXiv e-prints*, page arXiv:1705.10583, May 2017.

[13] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775):568–572, September 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 06 2016.

[15] Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes ImageNet good for transfer learning? *arXiv e-prints*, page arXiv:1608.08614, August 2016.

[16] Reto Knutti, Maria A. A. Rugenstein, and Gabriele C. Hegerl. Beyond equilibrium climate sensitivity. *Nature Geoscience*, 10(10):727–736, October 2017.

[17] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better? *arXiv e-prints*, page arXiv:1805.08974, May 2018.

[18] Andreas Kvammen, Kristoffer Wickstrøm, Derek McKay, and Noora Partamies. Auroral Image Classification With Deep Neural Networks. *Journal of Geophysical Research: Space Physics*, 125(10), 2020.

[19] J. Lee, W. Yoon, S. Kim, and et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2020.

[20] Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2016.

[21] V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.). IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. 2021.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[23] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Advances in Neural Information Processing Systems*, 2019.

[24] Stephan Rasp and Nils Thuerey. Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, 13(2):e2020MS002405, 2021.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. 05 2015.

[26] S. C. Sherwood, M. J. Webb, J. D. Annan, K. C. Armour, P. M. Forster, J. C. Hargreaves, G. Hegerl, S. A. Klein, K. D. Marvel, E. J. Rohling, M. Watanabe, T. Andrews, P. Braconnot, C. S. Bretherton, G. L. Foster, Z. Hausfather, A. S. von der Heydt, R. Knutti, T. Mauritsen, J. R. Norris, C. Proistosescu, M. Rugenstein, G. A. Schmidt, K. B. Tokarska, and M. D. Zelinka. An assessment of earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, 58(4), 2020. e2019RG000678 2019RG000678.

[27] Valentina Zantedeschi, Fabrizio Falasca, Alyson Douglas, Richard Strange, Matt Kusner, and Duncan Watson-Parris. Cumulo: A Dataset for Learning Cloud Classes. 2019.

[28] Jinglin Zhang, Pu Liu, Feng Zhang, and Qianqian Song. CloudNet: Ground-Based Cloud Classification With Deep Convolutional Neural Network. *Geophysical Research Letters*, 45(16):8665–8672, 2018.

[29] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL '21)*, 2021.

[30] Yanfei Zhong, Richen Ye, Tingting Liu, Zejun Hu, and Liangpei Zhang. Automatic Aurora Image Classification Framework Based on Deep Learning for Occurrence Distribution Analysis: A Case Study of All-Sky Image Data Sets From the Yellow River Station. *Journal of Geophysical Research: Space Physics*, 125(9), 2020.

## A Dataset Information

| Name | Subject | Task | Training Set Size | Test Set Size |
|---|---|---|---|---|
| ImageNet | Everyday Objects | Classification | 1,200,000 | 150,000 |
| CCSN | Clouds | Classification | 2,035 | 508 |
| SWIMCAT | Clouds | Classification | 628 | 156 |
| LGG | Brain MRIs | Segmentation | 6,680 | 1,178 |
| SWIMSEG | Clouds | Segmentation | 811 | 202 |
| SWINSEG | Clouds | Segmentation | 92 | 23 |
| SWINySEG | Clouds | Segmentation | 5,415 | 1,353 |
| Kiruna | Auroras | Classification | 3,000 | 846 |
| YR1 | Auroras | Classification | 1,080 | 120 |
| YR2 | Auroras | Classification | 7,201 | 800 |

Table 2: Datasets

## B Cloud Segmentation Results

| Source Dataset | Target Dataset | Train Accuracy (%) | Test Accuracy (%) |
|---|---|---|---|
| None | SWIMSEG | $88.63(\pm 0.00)$ | $84.80(\pm 0.01)$ |
| LGG | SWIMSEG | $88.96(\pm 0.12)$ | $83.29(\pm 0.44)$ |
| SWINSEG | SWIMSEG | $88.63(\pm 0.01)$ | $84.82(\pm 0.05)$ |
| SWINySEG | SWIMSEG | $93.69(\pm 0.01)$ | $\mathbf{86.60}(\pm\mathbf{0.05})$ |
| LGG → SWINSEG | SWIMSEG | $89.10(\pm 0.09)$ | $84.33(\pm 0.23)$ |
| LGG → SWINySEG | SWIMSEG | $91.43(\pm 0.00)$ | $85.60(\pm 0.01)$ |
| None | SWINSEG | $84.30(\pm 0.00)$ | $85.80(\pm 0.00)$ |
| LGG | SWINSEG | $85.81(\pm 0.02)$ | $87.00(\pm 0.05)$ |
| SWIMSEG | SWINSEG | $85.34(\pm 0.00)$ | $86.16(\pm 0.00)$ |
| SWINySEG | SWINSEG | $90.38(\pm 0.00)$ | $87.29(\pm 0.02)$ |
| LGG → SWIMSEG | SWINSEG | $86.73(\pm 0.00)$ | $86.62(\pm 0.01)$ |
| LGG → SWINySEG | SWINSEG | $88.15(\pm 0.00)$ | $\mathbf{87.48}(\pm\mathbf{0.00})$ |
| None | SWINySEG | $91.40(\pm 0.13)$ | $59.76(\pm 2.97)$ |
| LGG | SWINySEG | $90.77(\pm 0.07)$ | $\mathbf{65.68}(\pm\mathbf{8.38})$ |

Table 3: Cloud Segmentation Results (best result for each target dataset is bolded and one standard deviation is displayed in parentheses).

## C Aurora Classification Results

| Source Dataset | Target Dataset | Train Accuracy (%) | Test Accuracy (%) |
|---|---|---|---|
| None | Kiruna | $98.73(\pm 0.26)$ | $94.85(\pm 0.37)$ |
| ImageNet | Kiruna | $99.53(\pm 0.12)$ | $\mathbf{96.24}(\pm\mathbf{0.49})$ |
| YR1 | Kiruna | $99.03(\pm 0.13)$ | $94.07(\pm 0.46)$ |
| YR2 | Kiruna | $99.48(\pm 0.08)$ | $95.24(\pm 0.21)$ |
| ImageNet → YR2 | Kiruna | $99.82(\pm 0.09)$ | $95.66(\pm 0.48)$ |
| None | YR2 | $99.82(\pm 0.05)$ | $90.83(\pm 0.65)$ |
| ImageNet | YR2 | $99.86(\pm 0.05)$ | $\mathbf{93.85}(\pm\mathbf{0.55})$ |
| Kiruna | YR2 | $99.83(\pm 0.06)$ | $90.76(\pm 0.7)$ |
| YR1 | YR2 | $99.81(\pm 0.07)$ | $90.68(\pm 0.5)$ |
| ImageNet → Kiruna | YR2 | $99.90(\pm 0.04)$ | $93.81(\pm 0.48)$ |

Table 4: Aurora Classification Results (best result for each target dataset is bolded and one standard deviation is displayed in parentheses).