Authors' responds to Anonymous Referee 3:

We would like to thank the reviewer for the thorough review of our manuscript and insightful feedback. These comments have significantly improved the quality of our work. In the following sections, we present the reviewer's comments (in black), our responses (in red), and the changes made in the revised manuscript (in blue). Please note that all line numbers in our responses correspond to those in the revised manuscript.

Comments

1. The computational requirements and training times for both CSSL and baseline models should be discussed, as these are relevant for practical implementation. The community would highly benefit if the code and data for the CSSL algorithm were made publicly available.

   The data, code and models presented in this publication will be made available at: https://zenodo.org/communities/eth_zurich_iac_atmospheric_physics/

   The DOI links will be activated for public access upon acceptance of publication.

   In terms of the computational requirements, we stated the GPU information in Section. 4.1. We now add more detailed information about the hardware environment

   L273-L274

   Beyond the GPU requirements, the algorithm requires a minimum computing environment consisting of a 4-core CPU and 16GB of system memory to operate.

   In terms of the training times of our algorithm, the information is now concluded in the Appendix B.

   Appendix B

   The training times of models studied in this paper are listed in Table B1. The unsupervised pre-training phase required 66 minutes for the MoCo structure and 78 minutes for BYOL. For both the 19-category and 4-category classification tasks, the supervised fine-tuning phase of our models (Semisup-MoCo and Semisup-BYOL) consumed equivalent training time as their respective baseline models when trained on equivalent dataset sizes, hence, they are not displayed separately.

	Network	Size of training set (n)	Time (min)
Unsupervised pre-training	MoCo	33354	66
	BYOL	33354	78
Supervised fine-tuning/Baseline	ResNet50	128	5
		256	8
		512	13
		1024	29
		2048	38
		4096	44
		8192	66
		16384	85
		18864	96

Table 1: The training times of models.

2. Figure 1 would benefit from additional scale bars

The sample images shown in Figure 1 have been uniformly resized for demonstration purposes and do not reflect the actual physical dimensions of the ice crystals as captured by the imaging system. Therefore, it is hard to add scale bars while keeping the current figure unchanged. For detailed dimension information, Zhang et al. (2024) shows some sample images for each category with scale bars, which used the same dataset as this study.

L120-L121
A comprehensive collection of ice crystal examples can be found in the appendix of Zhang et al. (2024), where images of each distinct category are presented with scale bars indicating their actual dimensions.

3. Could you discuss the potential for transfer learning to other imaging systems that capture lower quality crystal images (e.g., those with coarser resolution) such as: VIZZZ, PIP/2DVD.

It would be highly interesting to test ice crystal images with a coarser resolution, keeping in mind that performance depends more on the number of pixels than on the actual size of the ice crystal. There are two potential approaches to transfer learning in this context. There are two potential approaches to transfer learning in this context. The first approach, as we mentioned in the conclusion section, involves incorporating the new dataset into our existing dataset and further training the pre-trained encoder within the upstream network, which can help us develop a foundational model for ice crystals that can be effectively transferred to downstream tasks. The second approach is also the main outcome of our research. It involves fine-tuning the pre-trained encoder directly on new datasets of varying sizes and categories. We now add some discussions about new imaging systems by following your comment.

L453-L456
…As the model has learnt the features of ice crystal in the upstream network, it can also adapt to data collected using different imaging devices such as VIZZZ (Maahn et al., 2024) and PIP&2DS (Jaffeux et al., 2022) with being fine-tuned on a small subset of new data, but the performance on such devices is required to be further evaluated in future studies…

L476-L382
…Expanding the scale of unsupervised pretraining enables the integration of datasets collected from different imaging probes, including (VIZZZ (Maahn et al., 2024) and PIP&2DS (Jaffeux et al., 2022)) possible…

4. Minor comments:
   Fixed, thanks

5. While rerunning the analysis with fewer convolutional layers would be beyond the scope of this review, it would be valuable if you could elaborate on the choice of network architecture and its implications. In particular, the use of 49 convolutional layers raises questions about computational efficiency versus model performance. Could a shallower network potentially achieve similar results with reduced computational overhead?

   The choice of 49 convolutional layers was based on our adoption of the ResNet-50 architecture, which was the common backbone across various computer vision tasks. A shallower encoder can improve computational efficiency while it may bring several concerns in the context of our task.

   In the upstream network, we require an encoder with sufficient capacity to extract the features of ice crystals to perform contrastive learning, especially for those complex shapes. A deep encoder can extract a hierarchy of features, from basic edges and textures in the shallow layers to complex shape patterns in deeper layers (Zeiler and Fergus, 2014). Therefore, a shallower ResNet may not be sufficient for the upstream network to extract the detailed information of ice crystals such as the complicated structures of aged particles or aggregates. We now add the above argument as a background information in the introduction section.

   L182-L185
   …A deep network can extract a hierarchy of features, from basic edges and textures in the shallow layers to complex shape patterns in deeper layers (Zeiler and Fergus, 2014). In our task of learning the features of ice crystals, it is necessary to a sufficiently deep network to extract the detailed information such as the complicated structures of aged particles or aggregates…

   Since the downstream network will directly use the encoder transferred from the upstream network, the architecture should keep the same as the encoder in the upstream network.

   In addition, as shown in the Appendix B, despite using the ResNet-50 architecture, the computational overhead for unsupervised pre-training remains practical and efficient. The supervised fine-tuning process also demonstrates reasonable computational demands, particularly when working with smaller subsets of around 2048 samples

   However, it is not necessary to conduct experiments with different depth of encoder. The choice of 49 convolutional layers in our implementation builds upon extensive prior research

on different visual tasks, demonstrating the effectiveness of ResNet50. In fact, the ResNet paper (He et al., 2016) had shown that ResNet-50, despite its greater depth, maintains comparable computational efficiency to the shallower ResNet-34 architecture. Specifically, ResNet-50 requires 3.8 GFLOPs compared to ResNet-34's 3.6 GFLOPs, while achieving a significant 3% reduction in error rate on ImageNet classification. We add a description of the effectiveness of ResNet50 in the manuscript.

L183-L185
…which was proved more efficient and effective than other variations of ResNet (He et al., 2016)…

Reference

Zeiler, M. D. and Fergus, R.: Visualizing and understanding convolutional networks, in: Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pp. 818-833, Springer, 2014.