Authors' responds to Referee Dr. Louis Jaffeux:

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. Public code and data:

   An associated GitHub repository or making the code and data public is highly encouraged. This article uses holographic imager data and could inspire researchers working with other image types, such as CCD imagers, optical array probes, or even 2D scattering probes, for which a wealth of hand-labeled datasets and trained algorithms already exist. The experiments could thus be easily reproduced with other data types and campaign datasets to validate the general conclusions on the CSSL algorithm.

   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.

2. Some improvements can be made in the presentation of each model in the tables. Initially, the semi-supervised models are not straightforward to identify in Tables 4 and 5, which may carry over into further reading of the study. The two unsupervised models are listed in Table 4, while the two semi-supervised models are labeled as supervised models (which is technically true). The fact that both tables do not directly correspond to the experiments, due to the inclusion of the "Unsup" models and the classification of CSSL-generated models as "Sup," may be confusing for readers unfamiliar with the employed technique.

   Thank you for pointing this out. Following your comments, we first clarify that the whole network is semi-supervised in the manuscript, and explain why we call it 'semi-supervised'

   Figure 2

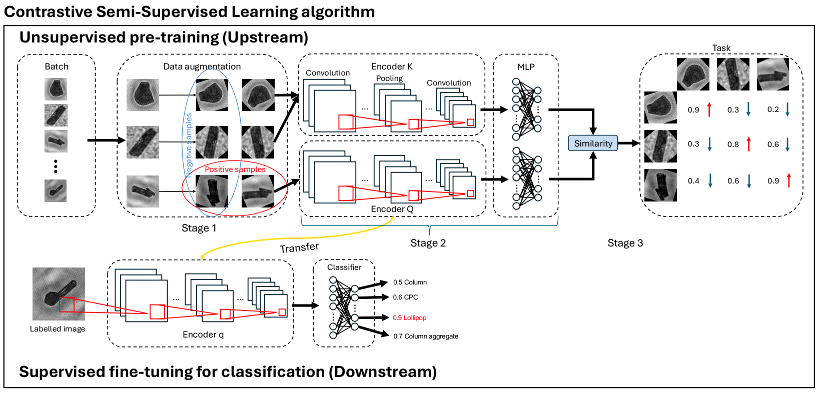

Figure 1: The schematic of CSSL algorithm

L131-L133

The downstream network adopts a traditional supervised image classification architecture, with a key distinction: the encoder is transferred from the upstream network. This transfer makes the whole network a semi-supervised learning approach, where the network uses both unsupervised knowledge from the upstream network and labeled image data from human.

L211-L213

The involvement of human knowledge (i.e. image labels) in the downstream network is the reason why we recognized the algorithm is semi-supervised.

We have renamed the models to reflect their types and structures. The type includes 'unsupervised (unsup)', 'semi-supervised (semisup)' and 'supervised (sup)'. 'Unsupervised' refers to the upstream network of CSSL, trained without labels. 'Semi-supervised' specifically refers to the classification stage of CSSL. 'Supervised' refers to purely supervised models. Structures include both MoCo and BYOL. In addition, we have renamed the weight initializations as 'dataset-type-structure' to clarify the source of the models' weights.

L275-277

In the rest of the paper, we will refer to a model by its type: unsupervised (unsup), supervised (sup) and semi-supervised (semisup); and its specific structure: MoCo and BYOL. The 'unsupervised' here represents the upstream network of CSSL algorithm. The 'semi-supervised' specifically refers to the classification stage of CSSL. 'Supervised' refers to purely supervised models.

L279-L280

the weights of both models are initialised with the weights of the respective structures pre-trained on the imagenet-1k dataset: IM1K-Unsup-MoCo and IM1K-Unsup-BYOL. In this paper, the weight initialisation will be refereed like 'dataset-type-structure'.

And then we modified the name of models and weight initialisations in Table 4 and 5 and in the manuscript according to our new naming systems. The unsupervised models are removed from tables since they are not the main object in the result section.

Table 4

Name	Dataset	Weight initialization	Size of training set (n)	Categories (c)
Semisup-MoCo	NASCENT19	NASCENT-Unsup-MoCo	[128, 256, 512, 1024, 2048, 4096, 8192, 16384, 18864]	19
Semisup-BYOL	NASCENT19	NASCENT-Unsup-BYOL	[128, 256, 512, 1024, 2048, 4096, 8192, 16384, 18864]	19
Sup (Baseline)	NASCENT19	IM1K-Unsup-MoCo	[128, 256, 512, 1024, 2048, 4096, 8192, 16384, 18864]	19
IceDetectNet	NASCENT19	IM1K-Sup	18864	19

Table 1: The models trained and used for studying the effect of training set size.

Table 4 caption

… "Semisup" represent the classification stage of CSSL which used both knowledge from unsupervised pre-training and image labels. "Sup" means that the models are purely supervised. "MoCo" and "BYOL" represent the encoders transferred from "Unsup-MoCo" or "Unsup-BYOL".

Table 5

Name	Dataset	Weight initialization	Size of training set (n)	Categories (c)
Semisup-MoCo-4CAT	NASCENT19-4CAT	NASCENT-Unsup-MoCo	[128, 256, 512, 1024, 2048, 4096, 8192]	4
Semisup-BYOL-4CAT	NASCENT19-4CAT	NASCENT-Unsup-MoCo	[128, 256, 512, 1024, 2048, 4096, 8192]	4
Sup-4CAT (Baseline-4CAT)	NASCENT19-4CAT	IM1K-Unsup-MoCo	[128, 256, 512, 1024, 2048, 4096, 8192]	4

Table 2: The models trained and used for studying the effect of number of categories.

3. Large error bars are found in Figures 4, 9, and 10 for small training sets. Additionally, the baseline model (fully supervised, with varying training set sizes) shows virtually the same performance as the two CSSL generated models. For the sake of transparency and setting realistic expectations for the paper, these limitations and the relative success of the experiments could be made more apparent in the abstract.

Thanks for your comments on the boxplot figures (Figure 4, 9 and 12). We did not have a clear and accurate elaboration of those results. We now improved our manuscript to explain the results more accurately. Firstly, we add explanations of each element in those boxplots. The central black lines in the boxes represent the median accuracy values of each 5-fold cross-validation experiment. The lower and upper limits of boxes are the 25% (the first) and 75% (the third) percentile (quantile) accuracy. Hence, the range of boxes displays the distribution of central 50% of accuracy values of each 5-fold cross-validation experiment, which means the average performance of each model. The error bars are the maximum and minimum accuracy values of each experiment, and the range of error bars shows the stability of each model.

Figure 4 caption

…The central black lines inside the boxes are the median values of accuracy of each 5-fold cross-validation experiment on the training set size. The lower limit and upper limit of boxes are the first quartile and the third quartile, respectively. The range of boxes shows the distribution of central 50\% accuracy values, which represents the average performance of each model. The error bars show the maximum and minimum accuracy values from each 5-fold cross-validation experiment…

After the additional explanations of boxplots, we would like to respond your comment points: The error bars range of CSSL is large when the training set sizes are small (especially for n=128 and n=256). We have already stated the finding and potential causes in L322, Section 5.1 for Figure 4. To make it clearer, we add more explanation:

L323-L329

One possible reason we concluded from checking the loss tendency during supervised fine-tuning on different sizes of dataset (Figure A2) is that the models fine-tuned on small sizes of dataset ($n< 2048$) is suboptimal compared to models fine-tuned on larger sizes, which would lead to unstable classification performance. Another possible reason we concluded from the loss value of unsupervised pre-training (Figure A1) is that the 33354 images may not be sufficient for optimizing the upstream network, which means the classification performance of CSSL algorithm could be further improved even when fine-tuning small size dataset if we pre-trained with more ice crystal images. We include the loss values in Appendix A.

We now add the same reason for Figure 9 in L388, Section 5.2.

L388-L389

The accuracy range of both the baseline and CSSL models trained on small datasets ($n < 2048$) are large. The reason is the same as the models fine-tuned on 19-category dataset.

In terms of Figure 10, it shows the trade-off between time saved on manual labeling and classification overall accuracy, which does not include the analysis of error bars. Therefore, we thought you may refer to Figure 12. As for Figure 12, we did not observe large error bar when the training size is small. At last, for the sake of transparency and setting realistic expectations, we add the content about large error bar in the abstract and the result section.

Abstract

L21-L23

Our analysis also reveals that both CSSL and purely supervised algorithms exhibit inherent instability when trained on small dataset sizes, …

Section 6

L439-L442

The second issue is that the classification performance of models fine-tuned on small dataset ($n < 2048$) is unstable. A possible reason is that the unsupervised model is not well pre-trained due to a relatively small size of unlabeled dataset (33354 samples). It could be solved by collecting more ice crystal images and feeding to the algorithm during unsupervised pretraining.

In order to prove our arguments about large error bar, we add figures of loss values during training in the Appendix A. For proving that 33354 images may not be sufficient for unsupervised pre-training, we draw the correspond loss values.

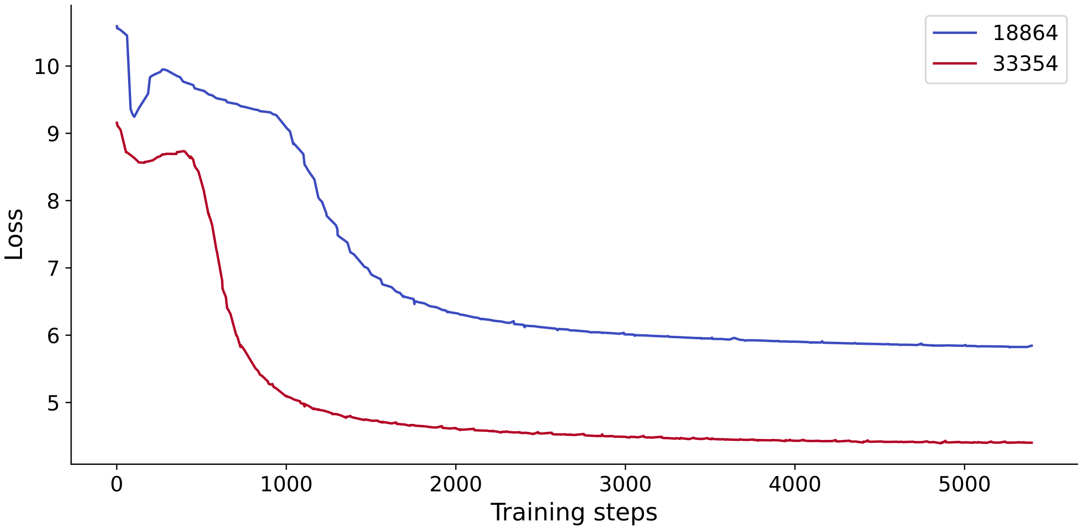

Figure 2: The loss of Unsup-MoCo during unsupervised pre-training use different size of unlabeled dataset

The legend means the number of unlabeled image samples we used for unsupervised pre-training. We tried to pre-train the network use only NASCENT19 dataset (18864 samples) and the whole NASCENT dataset (33354 samples). It can be found that the loss values of the model pre-trained on 33354 samples are completely lower than the one pre-trained on 18864 sample, which indicates the network trained on 33354 samples converged better. Therefore, it is reasonable to assume the classification performance of models would be improved and more stable if we involve more data (>33354) during the stage unsupervised pre-training.

The models' performance become unstable during supervised fine-tuning because the volume of samples is not enough for models to converge. If we compare the loss values of Semisup-MoCo during the supervised fine-tuning in the following figure, we can clearly find that at the last step, the loss value is much higher when the training set size are 128 and 256, which reflect the fact that the models trained with 128 and 256 samples are not well converged.

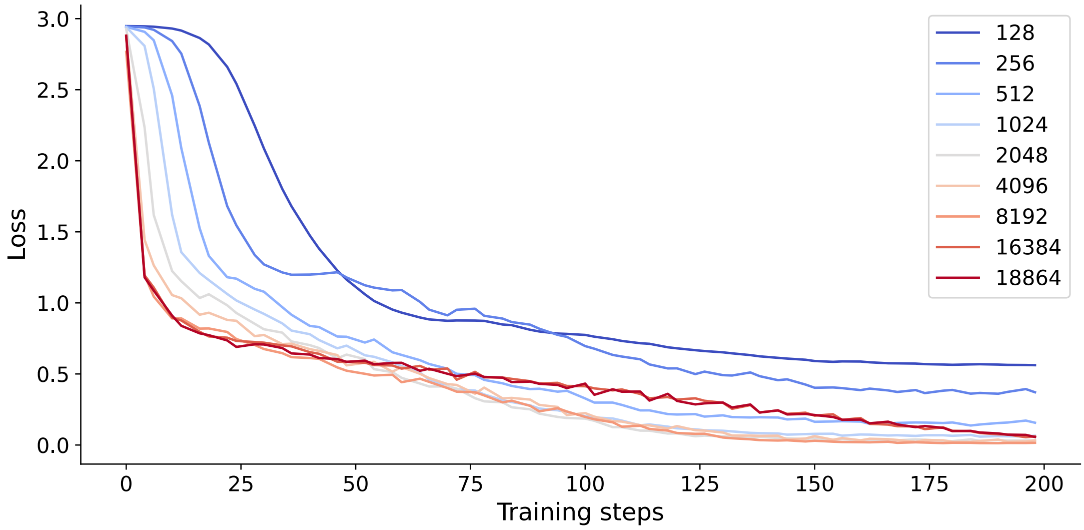

Figure 3: The loss of semisup-MoCo of different training set sizes.

Appendix A.

The loss value shows how well a deep learning model is trained. When the loss value stop decreasing significantly, the model can be considered as converged. The loss value of last step can indicate the model performance. In general, the smaller the value, the better the model can perform. To investigate why models trained on small datasets exhibit unstable performance, we analyzed the training loss for both Unsup-MoCo (Figure A1) and Semisup-MoCo (Figure A2).

We conducted unsupervised pre-training experiments using the NASCENT19 containing 18,864 samples and the complete NASCENT dataset containing 33,354 samples. It can be found that model pre-trained on NASCENT converged at a lower loss value compared to the model pre-trained on the smaller NASCENT19 dataset. It indicates that increasing the size of unlabeled data in unsupervised pre-training leads to more effective feature learning, which in turn would improve downstream classification performance.

When examining the loss trending of semisup-MoCo models during the supervised fine-tuning, we observed that models trained on small datasets (128 and 256 samples) converged at higher loss values compared to those trained on larger datasets. It indicates the downstream network fine-tuned with small sizes dataset is not optimal, which could lead to unstable performance.

In terms of the close performance between baseline models and two CSSL models, we agreed that it is true when the training set sizes are larger 2048, because it can be found from the Figure 4 and 9 that the classification performance of semisup-MoCo is obviously better than the baseline when the training set size was lower than 2048. It shows the strong classification performance of MoCo based CSSL when the training set size used for fine-tuning is small (n<2048). In fact, the average classification accuracy of all models converges to a similar value when the training set size became larger than 2048, which reveals the existence of a threshold beyond which increasing training set sizes yields diminishing returns in model classification performance improvement. We have already demonstrated it by Figure 5 and Figure 10 that there was a ‘inflection point’ for the trade-off between time spent on manual labelling and decreased overall accuracy.

Follow your advice, we add the statement of close performance among model when the training set size is large abstract for setting realistic expectations for the paper.

Abstract
L22-L23
…as well as the performance difference between them converges as the training set size exceeds 2048 samples…

4. Typos remain in the text. For example, in Section 3, the end of line 116, “useful fatures,” should be corrected to “useful features.”

Fixed, thanks!