2025 Joint ARM User Facility and ASR PI Meeting Rockville, Maryland.

Leveraging unlabeled ARM data with self-supervised pre-training for AI model optimization

Bhupendra Raut, Dario Dematties, Robert Jackson, Joe O'Brien, Max Grover, Seongha Park, Sean Shahkarami, Nicola Ferrier, and Scott Collis

Argonne National Laboratory

March 04, 2025

U.S. DEPARTMENT OF ENERGY

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Argonne NATIONAL LABORATORY

Why Machine Learning with ARM Data?

ARM's extensive observational network (radars, lidars, BL profiles, in-situ sensors) produces large volumes of data. Many remain unlabeled or partially labeled.

Challenges:

- Manual labeling for boundary-layer structures, cloud/rain features, is time-consuming.
- Existing supervised ML pipelines rely on annotated data, limiting the model's ability to discover novel features (e.g., anomalies in radar/lidar or synergy across instruments).
- Labeled subsets often represent only typical conditions, restricting learning of rarer events (e.g., boundary layer transitions, multi-layer cloud systems).

Need: A robust approach to use large-scale, unlabeled data from multiple platforms to improve model performance and reduce labeling costs.

Joint embedding architecture for self-supervised learning

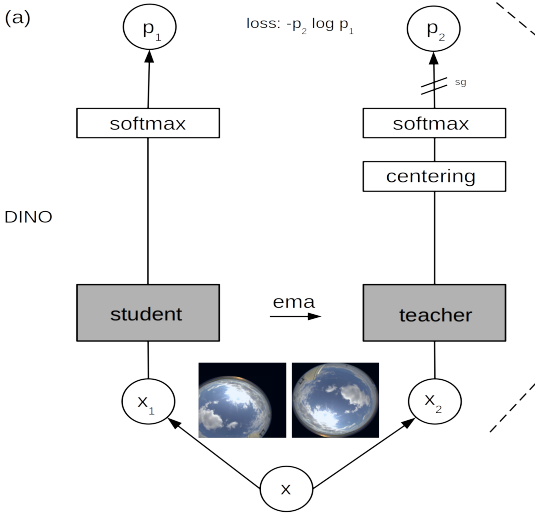
- SSL: compares augmentations without labels
- Focuses on embedding invariance and rich semantic features
- No direct reconstruction or discriminator modules, like GANs or AE

JEPA: Predictive approach with anchor vs. context; learns minimal, essential representations.
DINO: Teacher-student distillation without labels; enforces semantic alignment between different views.
VICReg: Three-term loss (Variance, Invariance, Covariance) to prevent collapsed embeddings.

Comparison with Autoencoders and GANs

- **Autoencoders:** Learn pixel-level reconstruction
- **GANs:** Adversarial generation through discriminator feedback
- **Joint Embedding:** Maximizes latent alignment, not reconstruction or generation

Joint embedding architecture for self-supervised learning

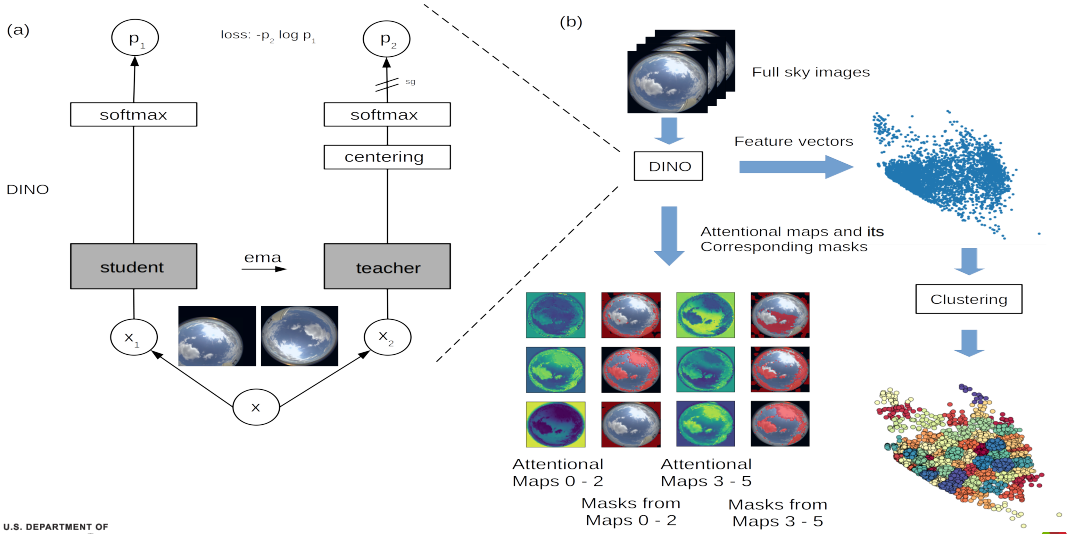


Distillation with no labels (DINO) (Caron et al., 2020, 2021)

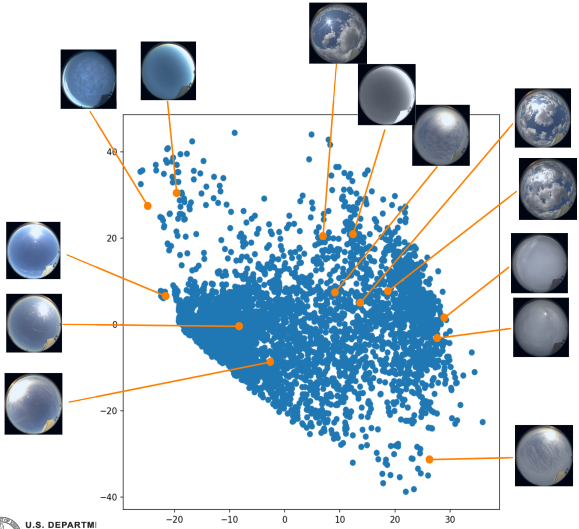
How it is intended to use?

1. Train with raw data first, then fine-tune with labeled data.
2. Any NN can be embedded with this architecture: ViT, CNNs.
3. Selectively ignore features (here, ‘color order’ and ‘values’).

Self-supervised learning for cloud segmentation and classification

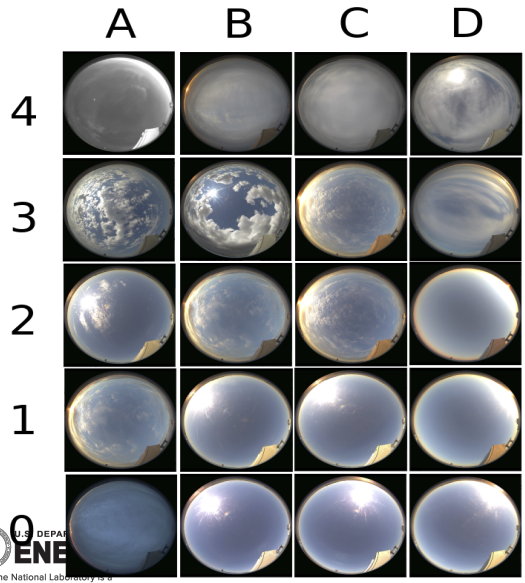


Self-supervised learning for classification (PCA1 Vs PCA2)



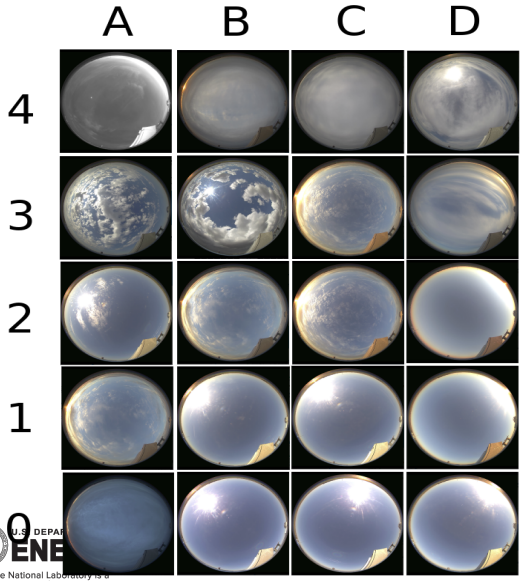
- Clear sky and cloudy images occupy two large regions.
- Overcast images with low and mid-level clouds separate into two regions.
- Partly cloudy images tend to split into multiple regions based on structure.

Self-supervised learning for classification (SOM)

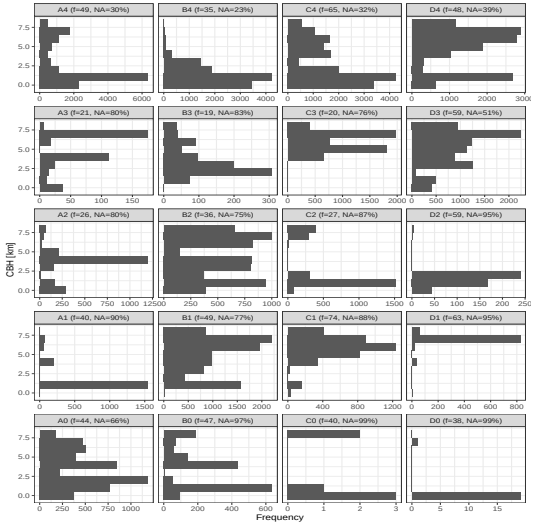


- Clusters reflect different cloud types, coverage, and diurnal peaks.
- More details in Dematties et al. (2023); Raut et al. (2023); Dematties et al. (2024)

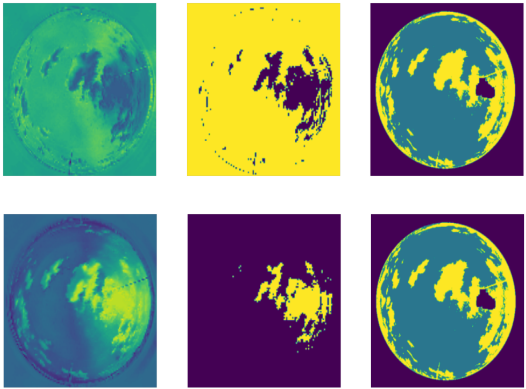
Self-supervised learning for classification (SOM)



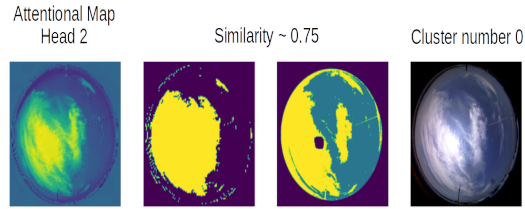
Clouds tend to form height-wise clusters (Dematties et al., 2023).



Attentional maps for segmentation without labels



- Normalized attention values yield cloud/no cloud segmentation.
- Attention aligns with cloud transparency.
- Note: Interpreting attention as transparency is nontrivial.



Method Overview: Unleash SSL on ARM Data

Key Idea: Train a model to learn general atmospheric representations from unlabeled data (radar, lidar, sky images, surface sensors), then fine-tune on small labeled datasets.

Workflow Steps:

1. **Raw Input Aggregation:** Collect diverse unlabeled data (e.g., Ka-Band ARM Zenith Radar reflectivity, MPL lidar backscatter, boundary-layer profiles).
2. **Augmentations:** Random transformations emphasize relevant variance (temporal shifts, intensity changes, 2D/3D subsetting).
3. **Joint Embedding SSL:** Contrastive or distillation-based methods (e.g., DINO, SimCLR) learn instrument-agnostic feature representations.
4. **Fine-Tuning:** Use minimal labeled sets (e.g., known cloud boundaries, aerosol types, or qc-labeled anomalies) to adapt the pretrained model.

Transfer Learning Potential: Pretrained embeddings accelerate downstream tasks (classification, retrieval, anomaly detection) for any new ARM campaign.

Foundation Models and Transfer Learning Outlook

- **Toward Foundation Models:** Train large-scale SSL across multiple ARM sites (SGP, NSA, ENA, mobile campaigns) to capture universal atmospheric patterns.
- **Cross-Platform Integration:** Incorporate scanning radars, aerosol measurements, surface flux towers, and more for a holistic Earth system representation.
- **Data Sharing and Collaboration:** Encourage community to pool unlabeled archives and share pretrained models to accelerate ML-based VAP development.
- **Uncertainty Quantification:** Combine SSL with Bayesian or ensemble methods to track predictive confidence for operational data quality insights.

Challenges and Practical Insights

- **Instrument Heterogeneity:** Different sampling rates, noise levels, scanning strategies; requires careful pre-processing.
- **Scalability and Compute:** SSL methods need large GPU resources for training on high-frequency ARM data streams.
- **Data Quality:** Varying QC levels across deployments; anomalies in spatiotemporal records can mislead SSL if not addressed.
- **Validation Gap:** Traditional metrics rely on labeled ground truth. SSL performance assessment needs alternative measures (clustering, geophysical consistency).
- **Community Adoption:** Many are cautious about black-box ML and seek stable, interpretable solutions. Clear guidelines for best practices are needed.

The biggest challenge is bridging the interdisciplinary divide between atmospheric scientists and AI researchers. Dedicated funding for collaborative AI exploration is essential.

References

Caron, M., I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, 2020: Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, **33**, 9912–9924.

Caron, M., H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, 2021: Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Dematties, D., S. Rajani, R. Sankaran, S. Shahkarami, B. Raut, S. Collis, P. Beckman, and N. Ferrier, 2024: Acoustic fingerprints in nature: A self-supervised learning approach for ecosystem activity monitoring. *Ecological Informatics*, **83**, 102 823.

Dematties, D., and Coauthors, 2023: Let's unleash the network judgement: A self-supervised approach for cloud image analysis. *Artificial Intelligence for the Earth Systems*, **00**, In 2nd review.

Raut, B. A., and Coauthors, 2023: A self-supervised approach for cloud image analysis. *103rd AMS Annual Meeting*, AMS.

Acknowledgements: This material is based upon work supported by the U.S. Department of Energy, Office of Science, under contract number DE-AC02-06CH11357. The Sage project is funded through the U.S. National Science Foundation's Mid-Scale Research Infrastructure program, NSF-OAC-1935984. The U.S. Department of Energy (DoE) Atmospheric Radiation Measurement (ARM) user facility supported the work under the field campaign AFC 07056 "ARMing the Edge: Demonstration of Edge Computing".