Supplementary information

Supp. Table 1: Collection of published work using automated methods for sea ice characterization based on remote sensing data.

Method	**Sensor**	**Channels**	**N Targets**	**Targets**	**Citation**
Dynamic Threshold	ERS-1	VV	3	FYI (2-classes), MYI	Fetterer et al. (1997)
Dynamic Threshold and Expert System	ERS-1	VV	3	Thin (< 30 cm), Medium (30 to 200 cm), Thick (> 200 cm)	Haverkamp et al. (1995)
Expert	ERS-1	VV	4	MYI, FYI rough, FYI smooth, YI	Holt et al. (1989)
SVM	RADARSAT-2	HH, HV (HH corrected for IA and HV noise corrected in pre-processing)	2	OW, Ice	Zakhvatkina et al. (2017)
RFC	Sentinel-1	GLCM from HH and HV	3	OW, FYI, MYI	Park et al. (2020)
CNN (3-layer)	Sentinel-1	HH, HV	4	OW, YI, FYI, MYI	Boulze et al. (2020)
CNN (VGG-16)	Sentinel-1	HH, HV, IA	2 and 5	OW, Ice; and OW, Brash, YI, level FYI,	Khaleghian et al. (2021)

				deformed + MYI.	
U-Net with attention	Sentinel-1	HH, HV, IA	2	OW, Ice	Ren et al. (2021)
U-Net (ensemble framework)	Sentinel-1	HH, HV	2	OW, Ice	Wang & Li (2021)
U-Net	Sentinel-1	HH, HV	11	Ice Concentration	Stokholm et al. (2022)
Autoencoder-decoder, Hierachical CNNs (AlexNet)	RADARSAT-2	HH, HV, mean(HH,HV)	5	OW, MYI, FYI, NI, YI	Chen et al. (2023)

Supp. Table 2: Percentage cover of *Oldest* Ice Type for each Extreme Earth v2 ice chart.

	New Ice	Nilas	Young Ice	First Year Ice	Old Ice	Water	Land
January	6.8	0.4	16.1	6.1	26.2	42.4	2.1
February	5.9	0.4	7.3	5.0	11.9	65.1	4.4
March	0.0	2.0	6.6	20.7	20.4	1.8	48.5
April	0.1	1.2	13.6	21.8	33.6	29.7	0.0
May	0.7	0.0	0.0	11.1	7.7	79.5	0.9
June	0.0	0.0	0.0	21.8	34.2	4.5	39.5
July	0.0	0.0	0.0	32.9	12.4	54.6	0.0
August	0.0	0.0	0.0	18.0	11.6	69.6	0.8
September	0.0	0.0	0.0	9.2	4.7	75.6	10.6
October	6.7	0.0	5.6	19.0	0.0	68.8	0.0
November	9.4	0.0	8.8	5.7	30.3	45.6	0.0
December	16.2	1.7	6.9	11.6	32.4	29.3	2.0
Total	4	0	5	15	19	47	9

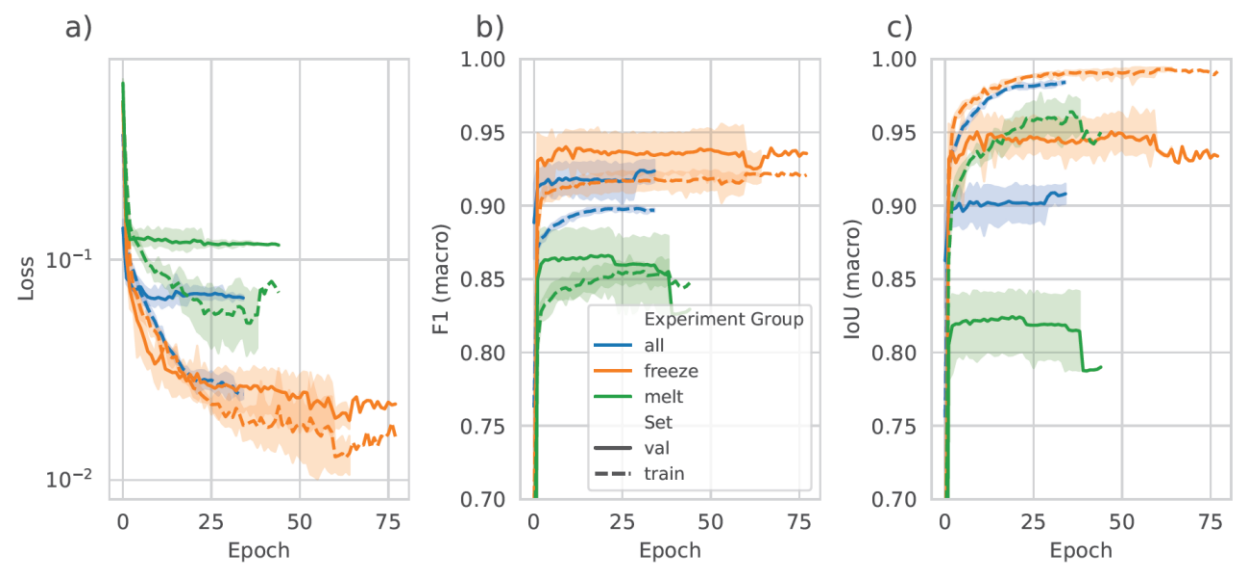
Supp. Table 3: Percentage cover of *Dominant* Ice Type for each Extreme Earth v2 ice chart.

	New Ice	Nilas	Young Ice	First Year Ice	Old Ice	Water	Land
January	17.9	0.5	8.1	23.8	5.4	42.4	2.1
February	6.7	3.6	5.4	14.0	0.8	65.1	4.5
March	0.4	3.2	6.9	29.1	10.1	1.8	48.5
April	0.1	2.0	26.6	35.5	6.2	29.7	0.0
May	0.7	0.0	0.0	17.5	1.4	79.5	0.9
June	0.0	0.0	0.0	48.2	7.3	5.1	39.5
July	0.0	0.0	0.0	40.8	4.6	54.6	0.0
August	0.0	0.0	0.0	24.9	4.7	69.6	0.8
September	0.0	0.0	0.0	11.7	2.1	75.6	10.6
October	9.4	0.0	7.9	14.0	0.0	68.8	0.0
November	9.8	1.9	9.2	31.4	2.1	45.6	0.0
December	16.2	4.2	14.8	30.3	3.1	29.3	2.0
Total	5	1	7	27	4	47	9

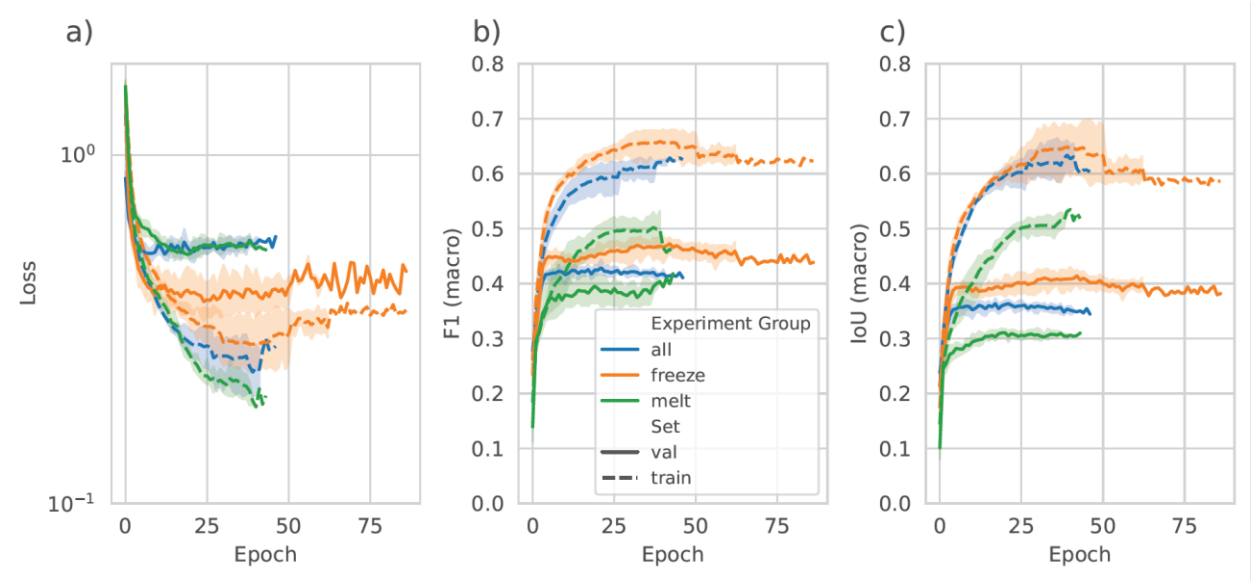
Supp. Table 4: Mean and standard deviation for each group. HH and HV values are in decibels, and incidence angle is measured in degrees. These statistics were calculated from training images listed in Table 3 in the main text for each experiment.

	Channel	all	freeze	melt
Mean	HH	-13.42	-11.72	-15.12
	HV	-27.01	-25.55	-28.46
	Incidence Angle	34.10	34.10	34.10
Standard Deviation	HH	5.42	4.80	6.04
	HV	5.10	5.12	5.08
	Incidence Angle	7.83	7.85	7.81

Supp. Figure 1: Binary training metrics. A) shows the training and validation losses during training, b) and c) show macro F1 and IoU, respectively. The three experiment groups are each composed of five different training executions with varying random seeds. The lines show the median value for the five runs, and the 95% confidence intervals are represented with a lighter background band. The models stop training if the validation loss does not decrease in 20 epochs, thus some models train for longer than others and the background aggregate is not displayed if there is only a single value for those epochs. During training, we choose to evaluate macro F1 and IoU that can better highlight low scoring classes for imbalanced datasets. The loss is generally smaller for the training set in comparison to the validation set. The figure also shows that training F1 and IoU generally do not show significant improvement for any experiment group after ~20 epochs, validation F1 and IoU seem to plateau even before that. In general, the loss charts confirm that the model is learning over epoch iterations and is also helpful to highlight possible overfitting.



Supp. Figure 2: Oldest ice type training metrics. A) shows the training and validation losses during training, b) and c) show macro F1 and IoU, respectively. The three experiment groups are represented by five different executions with varying random seeds. The lines show the median value for the five runs, the 95% confidence intervals are represented with a lighter background color. The models stop training if the validation loss does not decrease in 20 epochs, thus some models train for longer than others. In general, overfitting is more visible in the oldest ice segmentation experiments compared to ice/water discrimination, and results show a gap between the training and validation performance, with F1 and IoU difference close to 0.2 for the *all* and *freeze* experiment groups. While making general conclusions is difficult due to the small size of the dataset, this overfitting hints at model memorization, which may indicate the challenge and subjectivity of the ice types assigned in the label ice charts by the expert analyst.



Supp. Figure 3: Dominant ice type training metrics comparing models partially initialized with pre-trained weights against models with all weights randomly initialized. A) shows the training and validation losses during training, b) and c) show macro F1 and IoU, respectively. The two groups are represented by five different executions with varying random seeds. The lines show the median value for the five runs, the 95% confidence intervals are represented with a lighter background color. The models stop training if the validation loss does not decrease in 20 epochs, thus some models train for longer than others.