A. Model Architectures

A.1. UNet

UNet (Ronneberger et al 2015) is a CNN-based model that follows a symmetric encoder-decoder structure. The encoder consists of a series of convolutional layers, which are followed by ReLU non-linearities and pooling operations. This progressively reduces the spatial dimensions of the input image while increasing the depth of feature maps to capture relevant features at various scales. The decoder gradually reconstructs the original image size from the compressed feature representation by applying upsampling layers and convolutional layers. Essential for UNet are skip connections, which are concatenation operations that combine fine-grained feature maps from the encoder with semantically rich feature maps during the decoding process, thereby preventing loss of detail. The final layer is a 1×1 convolution that maps the feature maps to the number of output channels. The output layer is followed by a softmax activation function to convert logit outputs into class probabilities.

A.2. UNet++

UNet++ (Zhou et al 2018) introduces a series of nested, dense skip connections between encoder and decoder pathways, where encoder feature maps are passed through an additional dense convolution block before being combined with decoder feature maps. These skip connections gradually reduce the semantic gap between the encoder and decoder, resulting in more accurate results.

A.3. PSPNet

PSPNet (Zhao et al 2017) is another CNN-based method that addresses the limitation of fixed and local receptive fields. PSPNet incorporates multi-scale context information by using a spatial pyramid pooling module, which pools from the input feature map at different scales. Each pooled feature map undergoes a 1×1 convolution followed by upsampling. A final convolution is then applied to the fused features to make pixel-wise predictions.

A.4. DeepLabV3+

DeepLabV3+ (Chen et al 2018) is another multi-scale architecture that aims to combine contextual information with the ability to detect sharp boundaries, and builds on the problem that previous multi-scale architectures tend to lose detailed information. DeepLabV3+ uses dilated convolutions in the encoder, which allow the convolution operation to cover a larger receptive field (Chen 2017). It then adds a simple decoder that refines these features to recover spatial detail.

A.5. SwinTransformer

SwinTransfomer (Liu et al 2021) is a vision transformer-based architecture that uses shifted windows for efficient attention computation.

A.6. AutoSAM

AutoSAM (X Hu et al 2023) is an automated version of the transformer-based SAM model (Kirillov et al 2023). It uses a ViT encoder to process images as a sequence of patches by alternating multi-head attention layers and multi-layer perceptrons (MLPs). The encoded input features are further processed in an attention-based mask decoder, which maps the image embeddings to output masks. In contrast to the original SAM, AutoSAM does not necessitate input prompts (Kirillov et al 2023) for fine-tuning.

Table 1. *Validation results for different model architectures and pre-training datasets reported with respect to the sea ice class. Results are averaged based on 3-fold cross-validation.*

model (architecture - pre-training)	IoU (sea ice)	precision (sea ice)	recall (sea ice)	ROC-AUC (sea ice)
AutoSAM - none	0.739	0.843	0.832	0.857
AutoSAM - SA-1B	0.752	0.901	0.790	0.904
SAM2-UNet - none	0.732	0.900	0.754	0.868
SAM2-UNet - SAM2	0.746	0.923	0.760	0.889
UNet - none	0.723	0.870	0.763	0.858
UNet - ImageNet	0.765	0.907	0.793	0.876
UNet - RSD46-WHU	0.714	0.893	0.737	0.859
UNet - AID	0.762	0.908	0.787	0.882
PSPNet - none	0.656	0.894	0.694	0.832
PSPNet - ImageNet	0.639	0.918	0.657	0.832
PSPNet - RSD46-WHU	0.668	0.892	0.709	0.844
PSPNet - AID	0.685	0.926	0.711	0.859
DeepLabV3+ - none	0.629	0.907	0.656	0.825
DeepLabV3+ - ImageNet	0.637	0.852	0.681	0.819
DeepLabV3+ - RSD46-WHU	0.732	0.900	0.766	0.871
DeepLabV3+ - AID	0.730	0.886	0.772	0.871
UNet++ - none	0.738	0.899	0.764	0.859
UNet++ - ImageNet	0.754	0.911	0.783	0.859
UNet++ - RSD46-WHU	0.714	0.893	0.737	0.859
UNet++ - AID	0.764	0.91	0.785	0.876
SwinTransformer - none	0.416	0.648	0.659	0.481
SwinTransformer - ImageNet	0.766	0.909	0.787	0.886
SwinTransformer - Satlas	0.762	0.910	0.784	0.873

A.7. SAM2-UNet

SAM2-UNet (Xiong et al 2024) uses the multiscale Hiera encoder of SAM2 (Ravi et al 2024) and replaces its original transformer design with a UNet decoder containing convolutional and batch normalization layers.

B. Class-specific Performance

Sea ice and open water performance details are provided in Tables 1 and 2.

C. Class Balance Strategies

C.1. Patch Oversampling

We test strategies to address class imbalance and report results in Table 3. Patch oversampling redistributes the class frequencies in the training dataset by oversampling image patches that contain the minority class. To do so, 128×128 patches are sampled from the original 480×480 images that include melt pond pixels. The original 480×480 images are used for validation to make the results comparable to the baseline setup. The results show that performance decreases across all metrics. The decline may stem from the reduced spatial context available to the model and from the smaller effective training set due to unused image regions. This observation can be confirmed by training a baseline model on

Table 2. *Validation results for different model architectures and pre-training datasets reported with respect to the open water class. Results are averaged based on 3-fold cross-validation.*

model (architecture - pre-training)	IoU (open water)	precision (open water)	recall (open water)	ROC-AUC (open water)
AutoSAM - none	0.427	0.581	0.570	0.802
AutoSAM - SA-1B	0.582	0.659	0.719	0.885
SAM2-UNet - none	0.554	0.676	0.665	0.851
SAM2-UNet - SAM2	0.603	0.679	0.734	0.879
UNet - none	0.461	0.660	0.620	0.825
UNet - ImageNet	0.637	0.706	0.772	0.896
UNet - RSD46-WHU	0.57	0.593	0.778	0.868
UNet - AID	0.639	0.689	0.786	0.893
PSPNet - none	0.478	0.601	0.667	0.838
PSPNet - ImageNet	0.467	0.575	0.618	0.838
PSPNet - RSD46-WHU	0.506	0.677	0.618	0.841
PSPNet - AID	0.588	0.668	0.727	0.882
DeepLabV3+ - none	0.427	0.636	0.584	0.794
DeepLabV3+ - ImageNet	0.417	0.530	0.575	0.784
DeepLabV3+ - RSD46-WHU	0.578	0.708	0.682	0.881
DeepLabV3+ - AID	0.603	0.662	0.749	0.888
UNet++ - none	0.549	0.645	0.72	0.849
UNet++ - ImageNet	0.638	0.705	0.768	0.886
UNet++ - RSD46-WHU	0.57	0.593	0.787	0.874
UNet++ - AID	0.660	0.702	0.789	0.894
SwinTransformer - none	0.077	0.159	0.344	0.561
SwinTransformer - ImageNet	0.641	0.736	0.75	0.9
SwinTransformer - Satlas	0.635	0.712	0.763	0.893

(non-overlapping) size 128 patches. The further drop in performance of patch oversampling compared to the patch size 128 baseline may result from the fact that patch oversampling leads to the duplication of information in the already small dataset, as overlapping patches may be included.

C.2. Decision Rule

In the following, an approach inspired by Chan et al 2019 is tested, which applies class-wise weights to the output probabilities of the model before deciding for the final label. We modify the original method by computing priors on a class-wise rather than pixel-wise basis, as the latter assumption does not hold for overhead remote sensing imagery. As we observed overprediction of the melt pond class, we tested re-weighting by class frequency to make the melt pond class less likely to be predicted. The results indicate that precision can be significantly increased while the melt pond IoU remains similar and recall is lower (Table 3).

C.3. Loss Functions

Full loss function results are shown in Tables 4, 5, 6, 7, 8, and 9.

Table 3. *Test of different class balance strategies prior and post training. Patch sampling refers to oversampling training patches that include the minority class. Since the patch size used for this strategy (128) is smaller than the normal training input size (480), an ablation of training with patch size 128 without oversampling is reported. Decision rule refers to weighing output probabilities by class frequency of the training set. All results are reported for the best model (UNet++ pre-trained on AID) and averaged based on 3-fold cross-validation.*

model (balancing strategy)	IoU (melt pond)	IoU (sea ice)	IoU (open water)	precision (melt pond)	recall (melt pond)
decision rule	0.476	0.767	0.651	0.627	0.641
patch oversampling	0.362	0.627	0.432	0.451	0.670
baseline	0.486	0.764	0.660	0.563	0.760
baseline, ps 128	0.428	0.754	0.586	0.497	0.754

D. Hyperparameter Findings

In order to retrieve optimized hyperparameters, a 3-fold cross-validation hyperparameter search was performed with 100 runs across architecture - pre-training combinations. The following hyperparameters were explored: learning rate (0.0001, 0.0005, 0.001, or 0.01), batch size (1, 2, or 4), optimizer choice (SGD with 0.9 momentum, or Adam (Kingma and Ba 2017) with ($\beta 1$, $\beta 2$) = (0.5, 0.999)), and weight decay rate (0, 0.001, or 0.00001). The final hyperparameter settings were chosen based on melt pond IoU. For computationally demanding models (SAM2-UNet and SwinTransformer), only the learning rate was tuned, with the other parameters kept fixed at their previously identified optimal settings. Detailed per-model results are reported in the GitHub repository.

E. Semi-supervised Learning

In the following, the DMT algorithm (Feng et al 2022) is applied to the segmentation task. DMT is a pseudo-labeling method based on mutual learning between two models initialized with different weights. In these experiments, we employ UNet++ models pre-trained on AID and ImageNet as the base learners. Following the iterative training framework proposed in DMT, we used five iterations and a pseudo-labeled-to-labeled batch ratio of 7:1. The best-performing checkpoints were saved based on melt pond IoU. Training used a learning rate of 0.0005 and the Adam optimizer, based on the previous hyperparameter findings of this study. For pseudo-labeling and fine-tuning, 997 unlabeled images were used from the high-quality flights described in Section 2.2 of the main manuscript, ensuring that there is no overlap between them or with the labeled training images. This yields a labeled-to-unlabeled image ratio of 1:90.

Initial results show a modest improvement in performance (from 0.497 to 0.536 melt pond IoU), with minor fluctuations across iterations (Figure 1). Visual comparisons with the fully supervised baseline suggest a reduction in melt pond overprediction (Figure 2). These results demonstrate the potential of DMT to enhance model performance. Future work could involve optimizing the algorithm's hyperparameters and selecting more targeted subsets of unlabeled data for pseudo-labeling. The code is provided in the GitHub repository to allow for future adjustments (https://github.com/marlens123/tir_surface_segmentation).

Table 4. *Loss function findings for model architecture - pre-training combinations, reported with respect to the melt pond class.*

model (arch - pre-train - loss fn)	IoU (melt pond)	precision (melt pond)	recall (melt pond)	ROC-AUC (melt pond)
AutoSAM - none - Focal	0.192	0.274	0.401	0.736
AutoSAM - SA-1B - Focal	0.387	0.440	0.744	0.906
AutoSAM - none - Focal Dice	0.205	0.401	0.344	0.773
AutoSAM - SA-1B - Focal Dice	0.431	0.494	0.745	0.919
SAM2-UNet - none - Focal	0.353	0.405	0.706	0.869
SAM2-UNet - SAM2 - Focal	0.416	0.470	0.780	0.906
SAM2-UNet - none - Focal Dice	0.38	0.418	0.754	0.869
SAM2-UNet - SAM2 - Focal Dice	0.423	0.516	0.726	0.917
UNet - none - Focal	0.346	0.452	0.694	0.882
UNet - ImageNet - Focal	0.456	0.558	0.759	0.924
UNet - RSD46-WHU - Focal	0.431	0.509	0.761	0.912
UNet - AID - Focal	0.442	0.525	0.747	0.892
UNet - none - Focal Dice	0.279	0.45	0.543	0.869
UNet - ImageNet - Focal Dice	0.43	0.534	0.697	0.913
UNet - RSD46-WHU - Focal Dice	0.439	0.506	0.736	0.905
UNet - AID - Focal Dice	0.449	0.552	0.712	0.91
PSPNet- none - Focal	0.257	0.316	0.662	0.779
PSPNet- ImageNet - Focal	0.270	0.360	0.680	0.810
PSPNet - RSD46-WHU - Focal	0.320	0.358	0.683	0.831
PSPNet - AID - Focal	0.339	0.356	0.805	0.868
PSPNet - none - Focal Dice	0.195	0.361	0.413	0.752
PSPNet - ImageNet - Focal Dice	0.24	0.325	0.595	0.739
PSPNet - RSD46-WHU - Focal Dice	0.342	0.45	0.567	0.842
PSPNet - AID - Focal Dice	0.342	0.420	0.642	0.872
DeepLabV3+ - none - Focal	0.278	0.333	0.743	0.857
DeepLabV3+ - ImageNet - Focal	0.293	0.363	0.66	0.845
DeepLabV3+ - RSD46-WHU - Focal	0.427	0.443	0.797	0.891
DeepLabV3+ - AID - Focal	0.343	0.378	0.69	0.795
DeepLabV3+ - none - Focal Dice	0.292	0.327	0.616	0.830
DeepLabV3+ - ImageNet - Focal Dice	0.230	0.335	0.471	0.809
DeepLabV3+ - RSD46-WHU - Focal Dice	0.417	0.428	0.742	0.88
DeepLabV3+ - AID - Focal Dice	0.372	0.484	0.614	0.868

Table 5. *Continued.*

model (arch - pre-train - loss fn)	IoU (melt pond)	precision (melt pond)	recall (melt pond)	ROC-AUC (melt pond)
UNet++ - none - Focal	0.277	0.388	0.661	0.853
UNet++ - ImageNet - Focal	0.443	0.530	0.692	0.895
UNet++ - RSD46-WHU - Focal	0.449	0.511	0.804	0.927
UNet++ - AID - Focal	0.464	0.525	0.794	0.923
UNet++ - none - Focal Dice	0.244	0.451	0.503	0.859
UNet++ - ImageNet - Focal Dice	0.482	0.553	0.738	0.915
UNet++ - RSD46-WHU - Focal Dice	0.449	0.527	0.735	0.892
UNet++ - AID - Focal Dice	0.48	0.581	0.739	0.923
SwinTransformer - none - Focal	0.066	0.066	0.665	0.518
SwinTransformer - ImageNet - Focal	0.456	0.478	0.794	0.909
SwinTransformer - Satlas - Focal	0.469	0.509	0.77	0.903
SwinTransformer - none - Focal Dice	0.071	0.115	0.161	0.583
SwinTransformer - ImageNet - Focal Dice	0.468	0.514	0.778	0.922
SwinTransformer - Satlas - Focal Dice	0.463	0.526	0.725	0.898

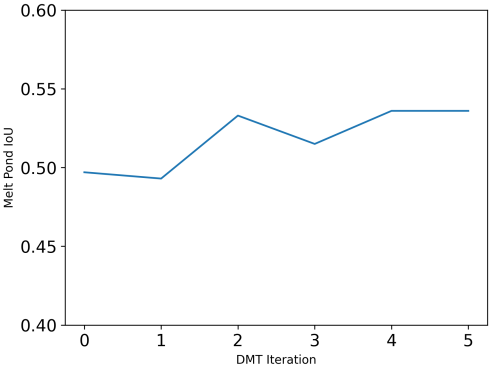


Figure 1. *Melt pond IoU plotted over DMT training iterations. Iteration 0 represents training on labeled data only (fully supervised). Note that the overall performance is slightly higher than that reported in the main results, since we report on a single validation set without k-fold cross-validation.*

Table 6. *Loss function findings for model architecture - pre-training combinations, reported with respect to the sea ice class.*

model (arch - pre-train - loss fn)	IoU (sea ice)	precision (sea ice)	recall (sea ice)	ROC-AUC (sea ice)
AutoSAM - none - Focal	0.724	0.826	0.834	0.840
AutoSAM - SA-1B - Focal	0.744	0.909	0.777	0.901
AutoSAM - none - Focal Dice	0.732	0.852	0.81	0.855
AutoSAM - SA-1B - Focal Dice	0.756	0.903	0.795	0.903
SAM2-UNet - none - Focal	0.721	0.904	0.74	0.858
SAM2-UNet - SAM2 - Focal	0.744	0.931	0.759	0.877
SAM2-UNet - none - Focal Dice	0.745	0.9	0.775	0.855
SAM2-UNet - SAM2 - Focal Dice	0.757	0.893	0.783	0.873
UNet - none - Focal	0.728	0.898	0.76	0.867
UNet - ImageNet - Focal	0.739	0.903	0.766	0.869
UNet - RSD46-WHU - Focal	0.745	0.905	0.768	0.877
UNet - AID - Focal	0.755	0.908	0.777	0.873
UNet - none - Focal Dice	0.751	0.87	0.802	0.857
UNet - ImageNet - Focal Dice	0.763	0.893	0.790	0.868
UNet - RSD46-WHU - Focal Dice	0.759	0.892	0.796	0.873
UNet - AID - Focal Dice	0.763	0.891	0.794	0.876
PSPNet- none - Focal	0.624	0.893	0.664	0.836
PSPNet- ImageNet - Focal	0.604	0.926	0.621	0.817
PSPNet - RSD46-WHU - Focal	0.666	0.896	0.699	0.844
PSPNet - AID - Focal	0.657	0.922	0.686	0.855
PSPNet - none - Focal Dice	0.556	0.836	0.604	0.790
PSPNet - ImageNet - Focal Dice	0.561	0.880	0.587	0.756
PSPNet - RSD46-WHU - Focal Dice	0.696	0.844	0.783	0.843
PSPNet - AID - Focal Dice	0.720	0.859	0.79	0.861
DeepLabV3+ - none - Focal	0.628	0.916	0.649	0.829
DeepLabV3+ - ImageNet - Focal	0.508	0.892	0.526	0.830
DeepLabV3+ - RSD46-WHU - Focal	0.731	0.913	0.757	0.873
DeepLabV3+ - AID - Focal	0.448	0.919	0.465	0.795
DeepLabV3+ - none - Focal Dice	0.666	0.878	0.708	0.824
DeepLabV3+ - ImageNet - Focal Dice	0.681	0.847	0.737	0.829
DeepLabV3+ - RSD46-WHU - Focal Dice	0.707	0.897	0.739	0.863
DeepLabV3+ - AID - Focal Dice	0.71	0.885	0.750	0.85

Table 7. *Continued.*

model (arch - pre-train - loss fn)	IoU (sea ice)	precision (sea ice)	recall (sea ice)	ROC-AUC (sea ice)
UNet++ - none - Focal	0.729	0.887	0.766	0.862
UNet++ - ImageNet - Focal	0.742	0.898	0.776	0.865
UNet++ - RSD46-WHU - Focal	0.733	0.918	0.75	0.879
UNet++ - AID - Focal	0.746	0.911	0.771	0.869
UNet++ - none - Focal Dice	0.732	0.851	0.793	0.859
UNet++ - ImageNet - Focal Dice	0.771	0.901	0.802	0.915
UNet++ - RSD46-WHU - Focal Dice	0.764	0.909	0.79	0.877
UNet++ - AID - Focal Dice	0.769	0.901	0.803	0.880
SwinTransformer - none - Focal	0.2	0.200	0.333	0.521
SwinTransformer - ImageNet - Focal	0.752	0.915	0.769	0.884
SwinTransformer - Satlas - Focal	0.754	0.927	0.770	0.866
SwinTransformer - none - Focal Dice	0.622	0.73	0.813	0.65
SwinTransformer - ImageNet - Focal Dice	0.767	0.901	0.793	0.875
SwinTransformer - Satlas - Focal Dice	0.636	0.9	0.803	0.870

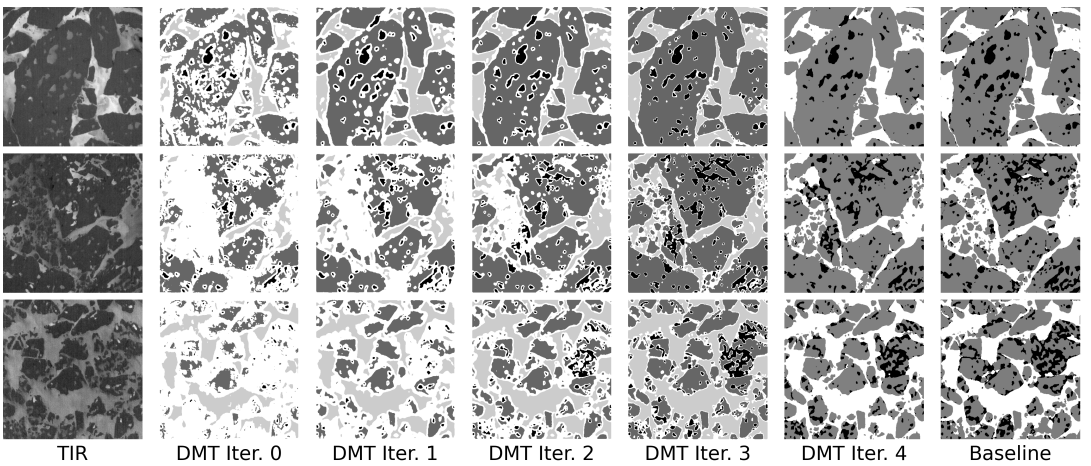


Figure 2. *Predictions from different DMT training iterations compared to the fully supervised baseline from the inference model. At each iteration, more top-confident pixels are pseudo labeled. White regions in the pseudo labels are ignored in the subsequent training step. We show pseudo labels from the AID-pre-trained model.*

Table 8. *Loss function findings for model architecture - pre-training combinations, reported with respect to the open water class.*

model (arch - pre-train - loss fn)	IoU (open water)	precision (open water)	recall (open water)	ROC-AUC (open water)
AutoSAM - none - Focal	0.301	0.567	0.4	0.77
AutoSAM - SA-1B - Focal	0.588	0.698	0.706	0.896
AutoSAM - none - Focal Dice	0.436	0.542	0.611	0.792
AutoSAM - SA-1B - Focal Dice	0.593	0.725	0.689	0.89
SAM2-UNet - none - Focal	0.502	0.673	0.625	0.841
SAM2-UNet - SAM2 - Focal	0.580	0.678	0.711	0.877
SAM2-UNet - none - Focal Dice	0.521	0.740	0.598	0.866
SAM2-UNet - SAM2 - Focal Dice	0.580	0.697	0.685	0.873
UNet - none - Focal	0.508	0.656	0.640	0.832
UNet - ImageNet - Focal	0.616	0.670	0.764	0.873
UNet - RSD46-WHU - Focal	0.613	0.672	0.750	0.88
UNet - AID - Focal	0.627	0.713	0.762	0.884
UNet - none - Focal Dice	0.449	0.618	0.585	0.836
UNet - ImageNet - Focal Dice	0.590	0.696	0.720	0.871
UNet - RSD46-WHU - Focal Dice	0.58	0.715	0.672	0.868
UNet - AID - Focal Dice	0.638	0.706	0.761	0.883
PSPNet- none - Focal	0.429	0.631	0.577	0.810
PSPNet- ImageNet - Focal	0.383	0.479	0.595	0.802
PSPNet - RSD46-WHU - Focal	0.529	0.651	0.674	0.845
PSPNet - AID - Focal	0.538	0.705	0.63	0.863
PSPNet - none - Focal Dice	0.377	0.519	0.615	0.792
PSPNet - ImageNet - Focal Dice	0.337	0.461	0.514	0.768
PSPNet - RSD46-WHU - Focal Dice	0.499	0.685	0.599	0.847
PSPNet - AID - Focal Dice	0.555	0.702	0.633	0.87
DeepLabV3+ - none - Focal	0.419	0.591	0.576	0.785
DeepLabV3+ - ImageNet - Focal	0.372	0.439	0.662	0.776
DeepLabV3+ - RSD46-WHU - Focal	0.597	0.732	0.718	0.872
DeepLabV3+ - AID - Focal	0.453	0.495	0.707	0.765
DeepLabV3+ - none - Focal Dice	0.445	0.675	0.581	0.821
DeepLabV3+ - ImageNet - Focal Dice	0.375	0.603	0.551	0.802
DeepLabV3+ - RSD46-WHU - Focal Dice	0.532	0.654	0.661	0.846
DeepLabV3+ - AID - Focal Dice	0.521	0.616	0.702	0.840

Table 9. *Continued.*

model (arch - pre-train - loss fn)	IoU (open water)	precision (open water)	recall (open water)	ROC-AUC (open water)
UNet++ - none - Focal	0.416	0.637	0.53	0.826
UNet++ - ImageNet - Focal	0.601	0.662	0.763	0.864
UNet++ - RSD46-WHU - Focal	0.614	0.681	0.755	0.876
UNet++ - AID - Focal	0.613	0.703	0.729	0.860
UNet++ - none - Focal Dice	0.363	0.633	0.494	0.8
UNet++ - ImageNet - Focal Dice	0.652	0.737	0.762	0.896
UNet++ - RSD46-WHU - Focal Dice	0.632	0.744	0.742	0.877
UNet++ - AID - Focal Dice	0.641	0.71	0.751	0.876
SwinTransformer - none - Focal	0.002	0.069	0.002	0.513
SwinTransformer - ImageNet - Focal	0.625	0.740	0.753	0.9
SwinTransformer - Satlas - Focal	0.642	0.715	0.769	0.89
SwinTransformer - none - Focal Dice	0.296	0.461	0.392	0.672
SwinTransformer - ImageNet - Focal Dice	0.63	0.745	0.724	0.892
SwinTransformer - Satlas - Focal Dice	0.636	0.756	0.724	0.893