**Methodology**

**Dataset**

The dataset used was recorded during ATWAICE Campaign 2022 in the central Arcitc. Available are 16 Flights with … images each, taken on consecutive days. Which camera and which spectrum. Flight height was 15m (ask Gunnar what to mention here, resolution). For training dataset creation, Flight 9 was picked due to good weather conditions. As only a few images could be labelled, this flight accorded for enough diversity and other flights were disregarded for this study. However, to accord for a more comprehensive representation, other flights, seasons and regions should be included in the training dataset.

The data is atmospherically corrected, however not projected to the ground surface yet. Implications are discussed in Section Discussion.

The files were retrieved from IUP server in netCDF format.

**Annotation**

Each pixels needs to be labelled in one of three categories: melt pond, sea ice, open water.

Out of flight 9, training images have been selected to accord for diversity (features with different size, shape and temperatures) and visibility to accord for accurate labelling.

Images have been upscaled to 2345,… and the resulting masks later downscaled. This might have caused some interpolation uncertainties.

For better visibility during the labelling process, temperature values have been clipped to 273, 276 which results in increased contrast. To simplify the labelling process, a preliminary mask has been created using thresholding and denoising (with scharr operator, which is an edge detection filter). The resulting outlines were used to create starting points for manual fine-tuning and filling. Gimp is used as annotation tool because it allows for high costumizability. Other labelling tools like LabelMe proved unsuccessful in first experiments (as no hard boundaries).

The labelled masks have been optimized and fine-tuned several times to accord for high accuracy. Corresponding VIS images were used. However, some parts are uncertain, see section Discussion.

In total, for one image, labelling process took several hours. Therefore, in total, only 8 images were labelled. Limiting factor in terms of training set size was labelling time, not data availability.

FIGURE: Show all training images used.

FIGURE: Scharr mask on image, intermediate labelling result.

**Data Preparation**

Training images and masks have been cropped to square, as Tensorflow is shown to perform best on this size and distortions at the border could be disregarded.

Masks have been one-hot encoded as standard for CNN.

Patch Extraction.

**Model Training**

“Lack of sufficient labelled data due to tedious and time-consuming manual segmentation”. Additionally, in the current task, ice floe and melt pond shapes have high intraclass variations due to different floe and pond sizes, temperatures…

With this limited data accessible for training… To successfully train a model with small data, overfitting is a big concern. This part tests the effectiveness of different methods that tackle this problem. To allow for comparibility, a baseline model was trained and successively methods added. The baseline model and methods are described below in more detail. Models have been evaluated using 5-crossfold validation as a small test set is likely to be biased. (How exactly is crossfold implemented). The main metric considered is mean IoU which is commonly used in segmentation problems. To save training time, all models have initially been trained for 40 epochs, as first experiments showed convergence at that point already. Configurations with slightly fluctuating training curves were trained again for 100 epochs to see if performance increases. This holds especially for augmentations, as this extends the training information available.

“For testing a model, patches of size K were extracted from the test image using a stride of K and segmentation was performed on each patch and the results were stitched back to get the final result.”

Baseline Model

* Optimizer is set to Adam which has an adaptive learning rate
* 0.2 was used as test ratio, resulting in 25 images for training and … for testing

**Evaluation Metrics**

Quantitative Segmentation Metrics:

1. Pixel Accuracy
2. Mean Intersection over Union (mIoU)
3. Per class IoU, as class frequencies are imbalanced
4. F1
5. (fw frequency weighted iou)

Melt Pond fraction (median of MAE over test set to be robust against outliers)

* Maybe mean IoU of unsupervised VIS and supervised

Qualitative Metrics: Compare 10 images (somehow subjective). “Additional results are in the supplementary material”. In case of smaller ice floes, water is misclassified as melt ponds several times…

**Discussion**

* Test set might be too similar to training set

**Transfer Learning: How transferable are features in deep neural networks?** (Yosinski) – Learning many related tasks at the same time with backpropagation (Caruana; 1995) – Deep lEarning of Representations for Unsupervised and Transfer Learning (Scotland 2012)

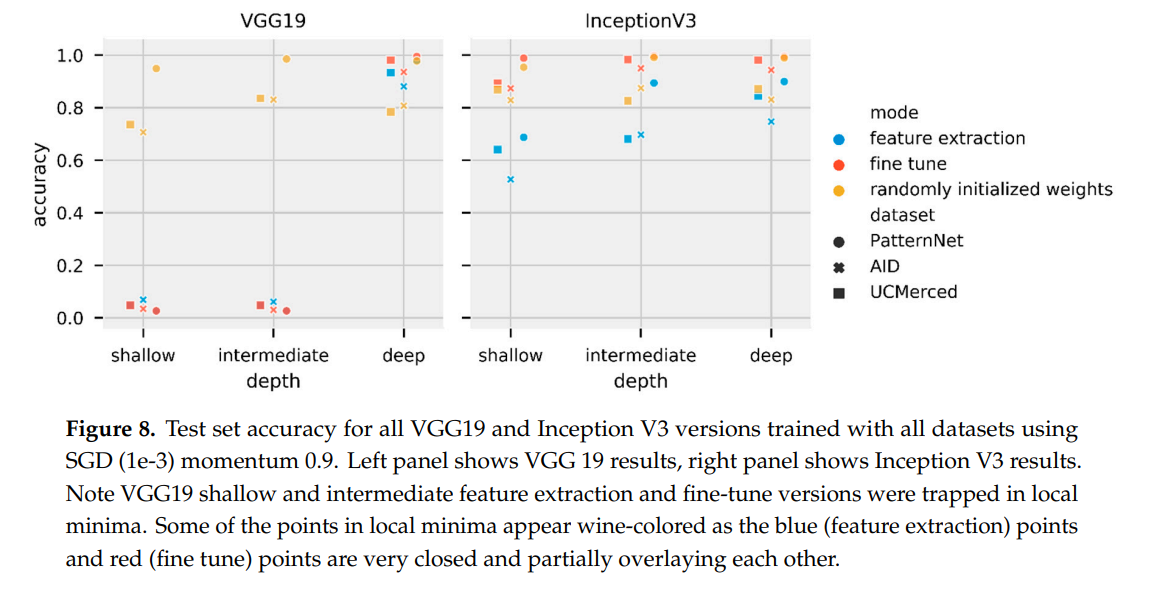
**Hu (2015): Transferring Deep CNN for the scene classificationof high-resolution RS imagery**

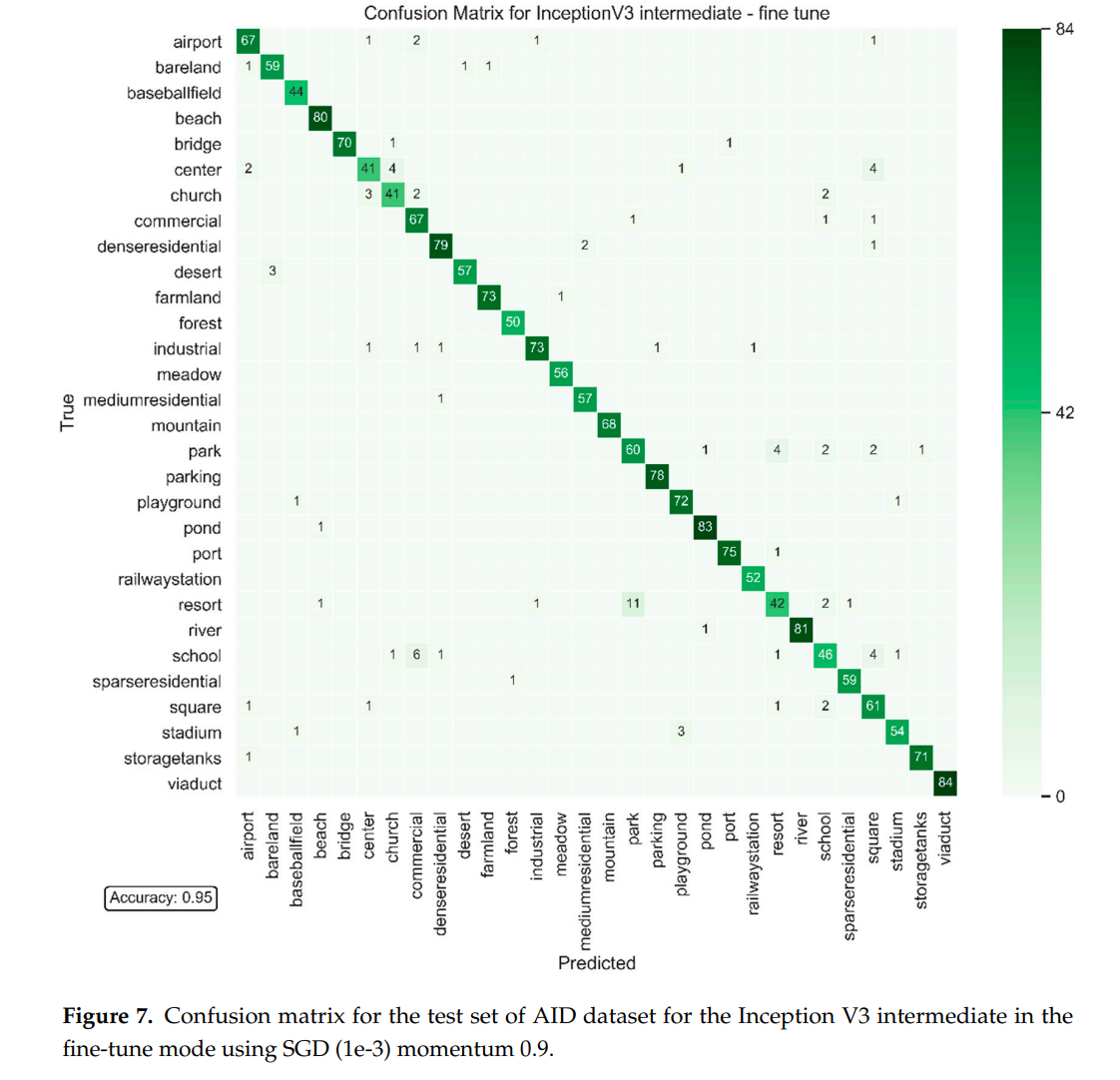
Yin (2017): Fine-tuning and visualization of CNN; Amsterdam

* The primary task (Imagenet, composed of natural images) is not very similar to the secondary task (RS). The number of channels differ, images are from different spectra

<https://www.mdpi.com/2072-4292/12/1/86> (Convolutional Neural Network for RS Scene Classification: Transfer Learning Analysis)

* “we find transfer learning from models trained on larger, more generic natural images datasets outperformed transfer learning from models trained directly on smaller remotely sensed datasets”
* Conventional scene classification techniques rely on low-level visual features to represent the images of interest. Such low-level features can be global or local. Global features are extracted from the entire remote-sensing image, such as color features, texture features and shape features. Local features are extracted from image patches that are centered about a point of interest. However, these global and local features are handcrafted, time-consuming, needs heuristic or ad hoc design decisions.
* Hu et al: the more representative and higher-level features, which are abstractions of the lower level features, are desirable and play a dominant role in scene classification task. The extraction of high-level features promises to be one of the main advantages of deep learning models.
* “Despite CNN’s powerful feature extraction capabilities, Hu et al found that in practice it is difficult to train CNNs with small datasets”
* Yosinski anf Yin observed that parameters learned by the layers in many CNN models trained on images exhibit a very common behaviour. The layers closer to the input data tend to learn general features, resulting in convolutional operators akin to edge detection filters, smoothing, or color filters. Then there is a transition to features more specific to the dataset on which the model is trained
* “In Transfer Learning, the filters learned by a CNN model on a primary task are applied to an unrelated secondary task. The primary model can be used as a feature extractor or as a starting point for a secondary model.”
* Despite the success of transfer learning in applications in which the secondary task is significantly different from the primary task (<https://www.nature.com/articles/nature21056>, <https://library.seg.org/doi/10.1190/segam2018-2998567.1> , <https://library.seg.org/doi/abs/10.1190/segam2019-3215401.1> ), the remark that the effectiveness of TL is expected to decline as the primary and secondary tasks become less similar (Yosinski) is commonly made. Although Yosinski concluded that using TL from distant tasks perform better than training CNN from scratch, it remains unclear how the amount of data or the model used can influence the performance.
* TL from natural images to RS is possible: Despite the relatively large difference between primary and secondary tasks, TL generally outperformed scratch. Fine-tuning models primarily trained on Imagenet outperformed finetuning on PatternNet, reason might be that Imagenet is complex dataset where intraclass variance (
* It is easier to overfit models with many weights
* When initial layer
* The backbone, when containing the weights learned during training for the primary task, will have its layers presenting the transition from general to specific features
* Explanation of SGD and why it is better than full gradient descent
* Three transfer learning modes: Feature extraction (freeze pre-trained layers), fine tuning (starts as feature extraction for half of the epochs but eventually allows all layers of the model to learn for half of the epochs), randomly initialized weights mode starts the entire model with randomly initialized weights after which all the weights are updated during training
* Validation and test set metrics better than training set metrics can be caused by the dropout layer, as during training less information is available for the model, or simply because of the data split; the training set is generally larger than validation and test sets and can incorporate a higher complexity in its samples
* Unlike poor generalization performance of adaptive methods compared to SGD optimizers reported by Wilson (65), results do not find significant differences in performance for optimizers tested. SGDs had slightly worse performance.
* Confusion matrix
* Provide information about training time (The executrion times are provided as simple general reference and lack more detailed analysis of performance – the computer was not entirely dedicated to experiments, thus speed might have been affected)
* In table, show the best performing training for each mode
* Adamax results in transfer learning in general better than SGD, step size 2e-3 too large for fine tune VGG19
* Difference in datasets: feature extraction mode limited. In fine-tuning overfitting reduces when layers are set to trainable.
* “Our results align with their findings”
* “all of the loss and accuracy per epoch can be accessed in the supplemental materials)
* Despite feature extraction limitations, the results show that transfer learning is an effective deep-learning approach that should not be discarded if the secondary task is not too similar to the primary task





<https://www.mdpi.com/journal/remotesensing/special_issues/DeepTransfer_Learning> (Special Issue)

<https://ieeexplore.ieee.org/document/8809071/references#references> (what data are needed for semantic segmentation in EO) few citations!

* Pretrained converges faster and better accuracy than from scratch

<https://ieeexplore.ieee.org/document/7301382/citations#citations> (Do deep features…)

* Evaluation of generalization power of pretrained ConvNets from everyday objects to the aerial and remote sensing domain
* “ConvNets have shown astounding results even in datasets with different characteristics from which they were trained, feeding the theory that deep features are able to generalize from one dataset to another.”
* “Although ConvNets were not the best descriptors for the coffee dataset, they could still perform well. This is interesting specially because the ConvNets used here were trained to recognize objects which is a very different scenario in relation to recognizing coffee regions. This also shows the generalization power of ConvNets”

<https://arxiv.org/pdf/1411.1792.pdf> (Yosinski: How transferable are features in deep neural networks?)

<https://ieeexplore.ieee.org/abstract/document/9122009>

used