MeltPondNet: A Swin Transformer U-Net for Detection of Melt Ponds on Arctic Sea Ice

Ivan Sudakow, Vijayan K. Asari, *Senior Member, IEEE*, Ruixu Liu, *Member, IEEE*, and Denis Demchev, *Member, IEEE*

***Abstract*—High-resolution aerial photographs of Arctic region are a great source for different sea ice feature recognition, which are crucial to validate, tune, and improve climate models. Melt ponds on the surface of melting Arctic sea ice are of particular interest as they are sensitive and valuable indicators and are proxy to the processes in the Arctic climate system. Manual analysis of this remote sensing data is extremely difficult and time-consuming due to the complex shapes and unpredictable boundaries of the melt ponds, and that leads to the necessity for automatizing the processes. In this study, we propose a robust and efficient automatic method for melt pond region segmentation and boundary extraction from high-resolution aerial photographs. The proposed algorithm is based on a swin transformer U-Net in which we introduce novel cross-channel attention mechanisms into the decoder design. The framework operates with optical data and allows for classifying imagery into four classes, i.e., sea ice/snow, open water, melt pond, and submerged ice. We use aerial photographs collected during the Healy–Oden Trans Arctic Expedition over Arctic sea ice in the summer season of 2005 as test data. The experimental results show that the proposed method is suitable for precise automatic extraction of melt pond geometry, and it can also be extended for other optical data sources that involve melt ponds. The approach has a promising potential to be used to analyze melt ponds' corresponding changes between years.**

***Index Terms*—Arctic, complex system, deep learning, melt ponds, remote sensing, sea ice, swin transformer.**

I. Introduction

COVERING 7%–10% of the the planet's surface, sea ice is a critical component of the Earth climate system and plays an important role in moderating global climate. In particular, sea ice moderates heat and gas exchange between the polar oceans and the atmosphere, reflects incoming solar radiation back into space, and serves as a home to marine life [1]. Temperature gradients between the Arctic and the lower latitudes also have a significant effect on atmospheric circulation patterns. Sea ice loss in the Arctic can be tied to the rapid warming trends observed recently in the Arctic, primarily due to the ice-albedo feedback [2].

In order to define the ice-albedo feedback, we must first define melt ponds. Melt ponds form atop Arctic sea ice during the spring/summer melt season from the melting snow layer on top built up over the winter months. Freshwater runoff from this snow melt begins to percolate through the porous micro structure of the ice, reducing the salinity of the brine in the pore space, causing it to freeze and block further drainage [3]. As the ice continues to warm, it becomes increasingly permeable eventually allowing the ponds to drain into the ocean below.

The ice-albedo feedback is the notion that as the ice begins to melt, ponds form atop the surface, lowering the albedo of the surface encouraging more melt, the melt further lowers the albedo and so on. This is an important effect to capture in any sea ice model being used for climatology and to date is not well-parameterized in climate scale sea ice models [4].

Not only do melt ponds have a significant effect on the energy budget of the Arctic, they also have an effect on the satellite-derived sea ice observations, in particular sea ice concentration (SIC) from passive microwave radiometry [5] for which the resolution is as low as 14–25 km, too low to resolve melt ponds. The high contrast in the microwave emissivity of sea ice is used to derive ice water fractions using tuned linear mixing models, which are taken as inputs representing satellite radiances at a variety of frequencies and polarizations. A major challenge in locating the melt ponds is that they have the same microwave signature as open water. In this way, they obscure the ice beneath them making it appear, as though there is less ice than those actually existing there. To combat this, the data derived "tie points" are used depending on the season. However, in cases where melt ponds are not present or in sufficient abundance, this can result in artificially inflated values of SIC. Errors can be as much as 30% [5]. The passive microwave record goes back almost 30 years and still can produce a daily snapshot of the ice cover giving a high-resolution long time record of the ice pack. Improving concentration retrievals during summer months is crucial for climate statistics, model evaluation, and for data assimilation with state-of-the-art models.

There is a long history of melt ponds defections on the images that include but not limited to the TerraSAR-X dual-polarization

Manuscript received 15 July 2022; revised 21 September 2022; accepted 6 October 2022. Date of publication 10 October 2022; date of current version 19 October 2022. This work was supported by the Division of Physics at U.S. National Science Foundation (NSF) under Grant PHY-2102906. *(Corresponding author: Denis Demchev.)*

Ivan Sudakow is with the School of Mathematics and Statistics, The Open University, MK7 6AA Milton Keynes, U.K. (e-mail: ivan.sudakow@open.ac.uk).

Vijayan K. Asari and Ruixu Liu are with the University of Dayton, Dayton, OH 45469 USA (e-mail: vasari1@udayton.edu; lrxjason@gmail.com).

Denis Demchev is with the Department of Space, Earth, and Environment, Chalmers University of Technology, 412 96 Gothenburg, Sweden (e-mail: denis.demchev@chalmers.se).

The dataset produced as a part of this study has been published and is publicly available for download at the link: https://zenodo.org/record/6602409.

Digital Object Identifier 10.1109/JSTARS.2022.3213192

data and airborne SAR images [6], [7], MODIS images [8], [9], the SHEBA and Healy–Oden Trans Arctic Expedition (HOTRAX) aerial photographs [10], seasonal sea ice monitoring and modeling site (SIMMS) field experiment photographs [11], and ENVISAT WSM images with HH-polarization [12].

The comprehensive analysis of the literature (see Appendices A and B) shows that the researchers prefer classical methods of image processing, ignoring machine learning approaches. It could be driven by different reasons, however the main one is that a method is chosen based on the "physics" of a solved problem. However, we need to use more universal methods of image analysis that would not rely on the set of certain physical parameters, but on the general principles of the developing system (geometry, complexity, physics, etc.). The main goal of this study is to develop a machine learning method for robust melt ponds detection that is based on general strategies of deep learning. CNN-based segmentation methods, such as the FCN [13], provide superior performance for natural image segmentation. The state-of-the-art models for image segmentation are variants of the encoder–decoder architecture, such as U-Net [14]. U-Net++[15] is essentially a deeply supervised encoder–decoder network where the encoder and decoder subnetworks are connected through a series of nested, dense, and skip pathways. With these hierarchical feature maps, the swin transformer model can conveniently leverage advanced techniques for dense prediction, such as feature pyramid networks FPN or U-Net. The transformer [16] is a network architecture originally developed for natural language processing (NLP). Also, inspired by the success of self-attention layers and transformer architectures in the NLP field, some works employ self-attention layers to replace some or all of the spatial convolution layers in the popular ResNet [17]. The visual transformer (ViT) [18] directly applies a transformer architecture on nonoverlapping medium-sized image patches for image classification. Swin transformer [19] modifies the ViT architecture to achieve the best speed–accuracy tradeoff among these methods on image classification. We also consider swin transformer as our main backbone and integrate it into the U-Net architecture with cross-channel attention, named as melt pond network (MeltPondNet), for melt pond detection on Arctic sea ice.

II. Data

In this study, we use aerial photographs of Arctic sea ice obtained during the HOTRAX captured from a helicopter between 5 August and 30 September, 2005 [20]. The flights have been typically flown at relatively low altitudes of 150–700 m to avoid the influence of low clouds with a digital camera Nikon D70 onboard. One thousand thirteen individual scenes over Arctic sea ice have been selected for the analysis, which contain highly detailed imagery of individual ice floes, melt ponds, and submerged ice and open water areas. The average photo resolution was 3042 $\times$ 2048 pixels. Depending on the altitude, the pixel resolution ranges from 5 to 25 cm per pixel. By visual expert analysis of the photographs, we defined zones with four classes of surface: 1) sea ice/snow; 2) melt ponds; 3) submerged ice; and 4) open water. These classes have been used for labeling in the training stage and as a set of model classes.

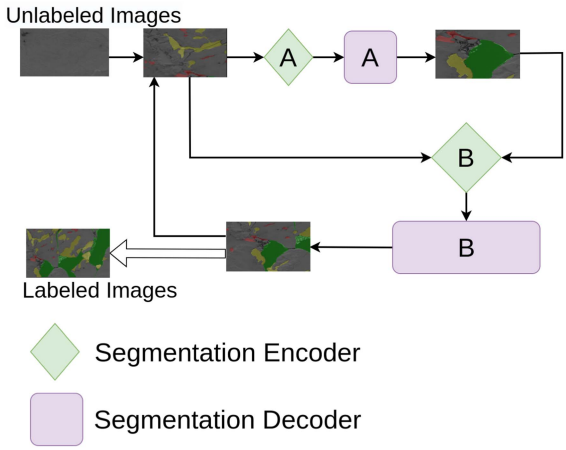


Fig. 1. Architecture of the CSSA method.

III. Methodology

A. Dataset Annotation

Many studies have focused on speeding up the image dataset annotation for semantic segmentation tasks. For example, one of the crowdsourcing methods is crowdsourcing annotations for visual object detection [21]. The three steps involved in this algorithm are: 1) drawing; 2) quality verification; and 3) coverage verification. In the drawing step, a worker draws one bounding box around one instance of the given image; in the quality verification step, a second worker verifies whether a bounding box is correctly drawn; and in the coverage verification step, a third worker verifies whether all object instances have bounding boxes.

Our cascaded annotation framework uses an incremental learning approach on a small batch of manually labeled images [22]. Then, it trains a segmentation model with the labeled data to propose segmentation areas on a batch of unlabeled images. Finally, it requests the annotator to correct possible incorrect polygons or label proposals. Thus, the involvement of human annotators is only in the correction stage [23].

Fig. 1 shows the conceptual diagram of our proposed method for cascaded semisupervised semantic segmentation annotation (CSSA). The segmentation encoder and decoder are trained on the dataset with unsupervised learning by reconstructing the input image [24]. Then, we changed the last layer of the decoder part to fine-tune the different classes in the dataset. The segmentation model is trained on a small set of manually annotated images. First, a trained model (i.e., model A) predicts pseudo labels on all unlabeled data. Next, model B is trained to annotate the unlabeled data by combining the labeled and pseudo-labeled data. After the first round of train-infer correction, the segmentation encoder and decoder parts are trained on

Algorithm 1: CSSA tool for image data labeling.

All images in the dataset to train a reconstruction encoder–decoder.
Set of all images in the dataset randomly splits to N+1 batches S_0, S_1, ... , S_N.
Set of pseudo labels P_{A_1}, P_{A_2}, ... , P_{A_N} created by models A_1, A_2, ... , A_N.
Set of suggested labels P_{B_1}, P_{B_2}, ... , P_{B_N} created by models B_1, B_2, ... , B_N.
L_0 ←manually annotate images in batch S_0.
for $i \in 1, 2, ..., N$ **do**
  if *part A* **then**
    model A_i ← train the segmentation network from the data S_0 to S_{i-1} with $L_0, L_1, ..., L_{i-1}$
    pseudo labels P_{A_i}–P_{A_N} ← predicted by model A_i
  else
    model B_i ← train the segmentation network with the data from S_0 to S_N with the labels $L_0, L_1, ..., L_{i-1}$ and $P_{A_i}, P_{A_{i+1}}, ..., P_{A_N}$
    suggested labels P_{B_i}–P_{B_N} ← predicted by model B_i
  end
  L_i ← do manual correction for the suggested labels
end
Fully labeled dataset L_0, L_1, ... , L_N.

the recently labeled batch. This process continues in a loop until all unlabeled batches are labeled.

The functional framework of the CSSA method uses unsupervised learning to obtain a feature encoder. Then, the CSSA model uses an incremental learning approach on a small batch of manually labeled images [25]. After that, we train a segmentation model with the labeled data to propose bounding boxes on a batch of unlabeled images and request the annotator to correct possible incorrect polygons or label proposals. In this process, the involvement of human annotators is only in the correction stage. Hence, our method decreases the tedious task of manual annotations. Algorithm 1, shown as follows, summarizes all the relevant steps of the proposed iterative training method.

The first step in the CSSA procedure is unsupervised training of the whole dataset to obtain a suitable encoder. The second step is fully annotate (manually) an initial batch of images from the unlabeled dataset. This stage is manual and requires human involvement to draw polygons and provide class labels on images. In this stage, we use a basic segmentation annotation tool (i.e., Labelme) with no extra speedup procedures to create mask labels. The third step is to train segmentation model A (supervised training) with the fully annotated data (i.e., *L*). Although any segmentation network can be used for this purpose, we focus on a recent deep learning-based semantic segmentation model of U-net [14], [26]. The fourth step is to train human-annotated initially labeled data and pseudo-labeled data together and relabel the pseudo-labeled data again. Now, the system outputs the human-annotated labeled data. Finally, the semisupervised model suggests labeled data (after predicting the unlabeled data by the network B). Before the cascaded network starts outputting the fully annotated data, the human annotator needs to correct the mask polygons suggested by model B and the annotated data shown in Fig. 2.

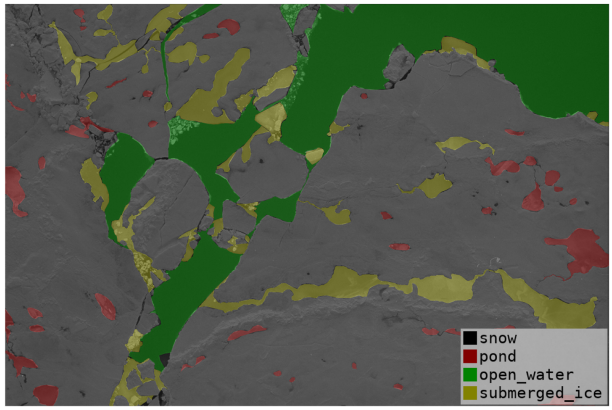


Fig. 2. Melt ponds data visualization.

Given an image $x \in H \times W \times 3$ with a spatial resolution of $H \times W$ and 3 channels (RGB). The network is to predict the corresponding pixel-wise target map with size H × W. The normal deep neural network is to directly train a U-Net, which first encodes images into high-level feature representations, and then decoded back to the full spatial resolution. Unlike existing approaches, our method introduces self-attention mechanisms into the encoder design [27] via the usage of transformers [19]. We will first introduce how to directly apply a transformer for encoding feature representations from decomposed image patches [28], [29]. The elaborated framework of the overall architecture is shown in Fig. 3.

Transformers take image into nonoverlapping patches by a patch partition module [18], [30]. Each patch is treated as a "token" and its feature is set as a sequence of vector. The self-attention mechanism in transformers projects each feature X into corresponding query, key, and value vectors, using learned linear transformations W^Q, W^K, and W^V. Thus, the projection of the whole sequence generates representations *Q*, *K*, and *V*, which is formulated as

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (1)$$

where d is the query or key dimension and the values in B are taken from a smaller sized bias matrix. The basic unit of MeltPondNet is swin transformer block [19]. We use it to substitute the traditional convolution layer in the U-Net module. The number of swin transformer layers is always a multiple of two where: 1) one is for window multihead self-attention (W-MSA); and 2) the other is for shifted W-MSA (SW-MSA). With the shifted window partitioning approach, consecutive swin transformer blocks are computed as

$$\begin{aligned} \hat{z}^l &= \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \\ z^l &= \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l \\ \hat{z}^{l+1} &= \text{SW-MSA}(\text{LN}(z^l)) + z^l \\ z^{l+1} &= \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \quad (2)$$

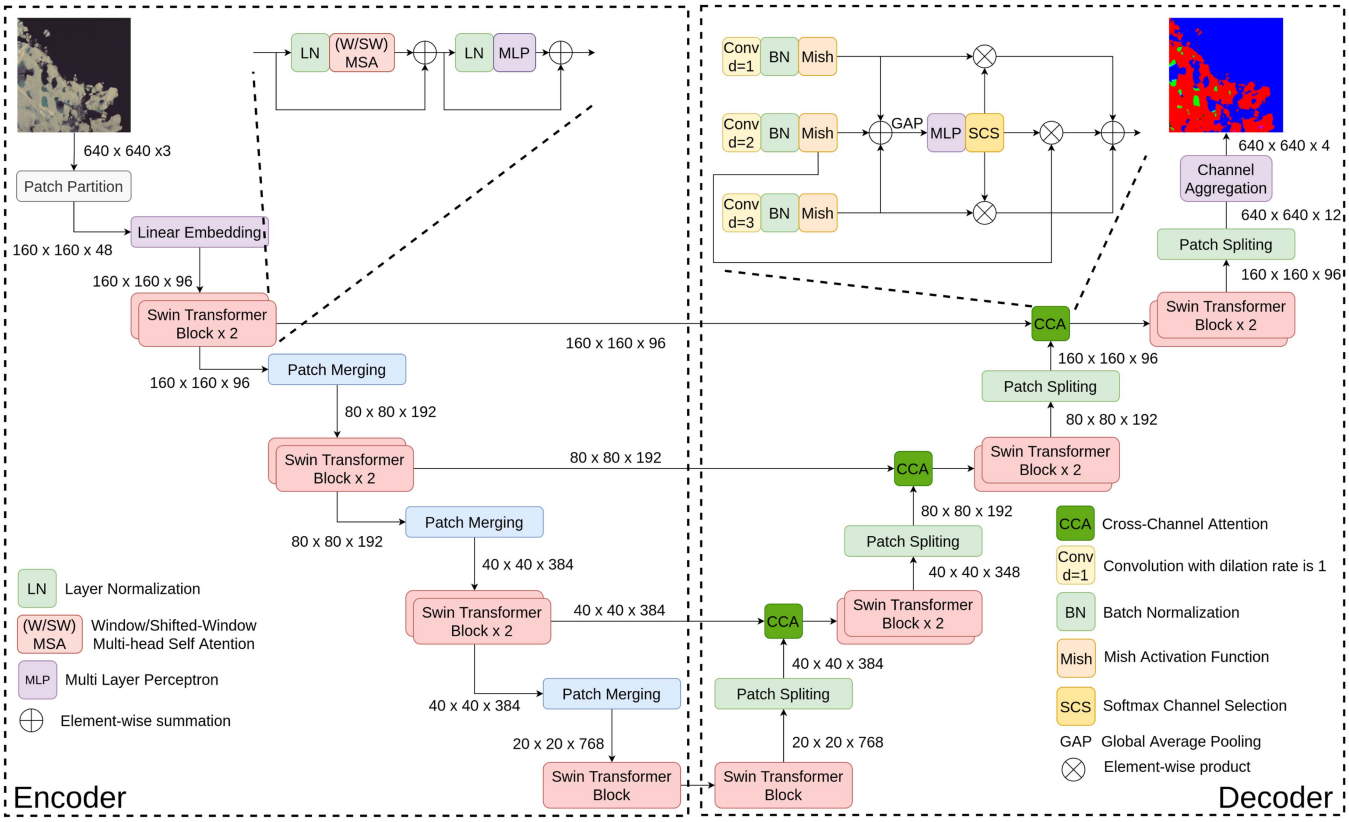


Fig. 3. MeltPondNet architecture that consists of an encoder, decoder, skip connections, and cross-channel attention to fuse the multiscale features.

where *LN*( ) denotes the layer normalization and MLP is a multilayer perceptron that has two fully connected layers with Gaussian error linear unit activation function.

MeltPondNet feeds the inputs into sequence embeddings for the encoder, and the geospatial images split into nonoverlapping patches with a patch size of 4×4. Furthermore, a linear embedding layer is applied to the projected feature dimension into an arbitrary dimension (we used 96 in this study). The patch-merging layer is the same as the original swin transformer structure. The input patches are divided into four parts and concatenated together by the patch-merging layer. With such processing, the feature resolution will be downsampled by two times. And, since the concatenate operation results in the feature dimension increase by four times, a linear layer is applied to the concatenated features to unify the feature dimension to the two times of the original dimension. Corresponding to the encoder, the symmetric decoder is built based on the swin transformer block. To this end, in contrast to the patch-merging layer used in the encoder, we use the patch-expanding layer in the decoder to upsample the extracted deep features. The patch-expanding layer reshapes the feature maps of adjacent dimensions into a higher resolution feature map (two times the upsampling) and reduces the feature dimension to half of the original dimension accordingly. The cross-channel attention module consists of three parallel operations: 1) dilated convolution; 2) batch normalization; and 3) Mish activation. It selects the important channel using different kernel sizes implemented by dilation convolution rates. We use depthwise separable convolutions to replace the standard convolution to save parameters and speed up the processing time. The dilated convolutions in the three parallel branches have the same kernel size but different dilation rates. Specifically, the kernel of each dilated convolution is 3×3, and the dilation rates d are 1, 2, and 3 for different branches. Dilated convolutions support exponentially expanding receptive fields without losing resolution or coverage. However, in the convolution operation of dilated convolution, the elements of the convolution kernel are spaced, and the size of the space depends on the dilation rates, which is different from the elements of the convolution kernel that are all adjacent in the standard convolution operation. The dilation rates 1, 2, and 3 dilation convolutions are approximately equal to kernel sizes 3×3, 5×5, and 7×7 standard convolution, respectively.

IV. Results and Discussion

The deep learning architectures U-Net [14], U-Net++[15], transformer U-Net [31], and the proposed MeltPondNet are implemented based on Python 3.8 and Pytorch 1.9. For all training cases, data augmentations, such as flips and rotations, are used to increase data diversity. We train our model on 4 NVIDIA TITAN RTX GPU with 24 GB memory. The synaptic weights pr-trained on ImageNet are used to initialize the model parameters. During the training period, the batch size is 8 and the popular SGD optimizer with momentum 0.9 and weight decay 1e-4 is used to optimize our model for error back propagation learning.

TABLE I
ABLATION STUDY ON THE DATASET

Data	Model	DSC	mIOU
Three classes	U-Net	93.12	88.01
	U-Net++	93.34	88.37
	Transform U-Net	93.82	88.92
	MeltPondNet	94.02	89.07
Four classes	U-Net	87.57	79.24
	U-Net++	88.02	79.82
	Transform U-Net	89.10	81.07
	MeltPondNet	89.50	81.46

TABLE II
EVALUATION MATRIX FOR DIFFERENT CATEGORIES

	Class	DSC	mIOU	F1	Accuracy
Transformer U-Net	Snow	95.88	93.48	95.54	97.15
	Pond	74.21	69.12	76.77	97.24
	Open water	71.35	78.31	82.41	98.68
	Submerged ice	48.07	58.60	64.44	97.17
MeltPondNet	Snow	95.63	94.59	96.05	97.95
	Pond	75.55	69.14	76.92	97.71
	Open water	72.74	78.47	82.61	98.98
	Submerged ice	49.28	58.81	64.93	97.27

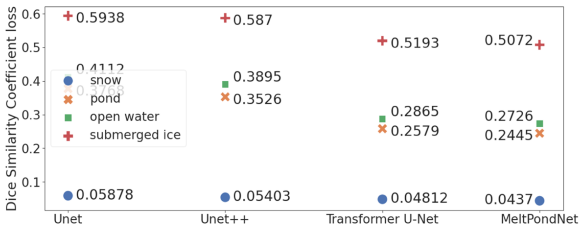


Fig. 4. Comparison of segmentation performance in terms of DSC loss (1-DSC/100) (the lower the better).

1) Model Performance Evaluation: We have two kinds of label strategies: a) one is three classes; and b) the other is four classes. The three classes are: a) snow; b) pond; and c) open water, and the fourth class is submerged ice. In Table I, we compare the performance of the U-Net, U-Net++, transformer U-Net, and MeltPondNet. It can be seen that the MeltPondNet shows the best overall performance for both label strategies, and the details for each class are given in Table II. As quality metrics, we use dice similarity coefficient (DSC) that combines the advantages of precision and recall, and mIOU, that is, the mean value of IoUs (a number from 0 to 1 that specifies the amount of overlap between the predicted and ground truth bounding box), corresponding to different classes which would match with the actual degree of similarity. F1 score is the harmonic mean of the precision and recall.

We can conclude from the obtained results that using the three classes provide a more robust classification by all models. Probably, it is caused by difficulties for a model to detect submerged ice because of its more complex and varying signature that consequently led to ambiguities with other classes. The details of the each class quantitative results are shown in Figs. 4 and 5

2) Robustness to Different Resolutions: Since our MeltPondNet has the fixed input image resolution, we preprocess the

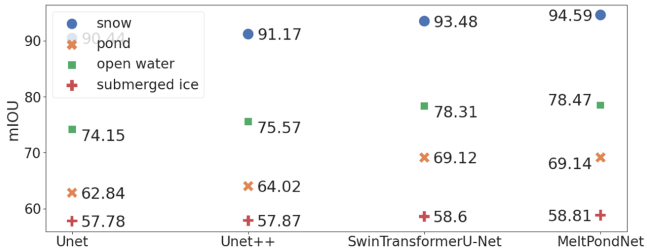


Fig. 5. Comparison of segmentation performance in terms of mIOU (the higher the better).

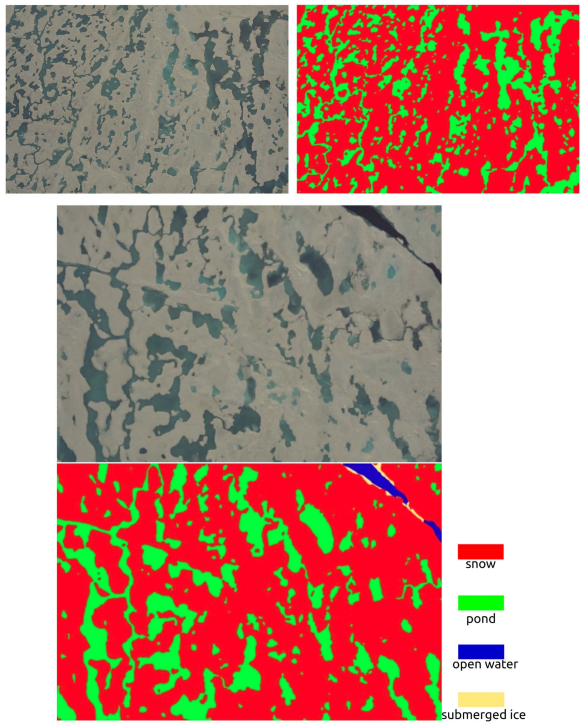


Fig. 6. MeltPondNet segmentation results for different resolutions.

input images by partitioning to many overlapping patches with 640×640 pixels. Then, segmentation forward pass is applied independently to each overlapping patch. Finally, the overlapping prediction results are merged back into the original size by weighted average. Based on the slicing patch method, the input image can be of any size. In Fig. 6, an example of a high-resolution image is shown, and that demonstrates the benefits of the high-resolution in reducing confusion with the mixed pixels. The result legend is shown at the right-hand side of the predicted results.

3) Robustness to Different Background: As shown in Fig. 7, we picked some different background results from our test dataset. Our MeltPondNet can detect the snow, pond, and open water classes very well, no matter how the environment is illuminated bright or dark. The submerged ice class is challenging to identify because it usually has a similar color to open water or the melt pond.

4) Robustness to Image Artifacts: Due to the melt pond images are usually acquired from the inclement weather area,

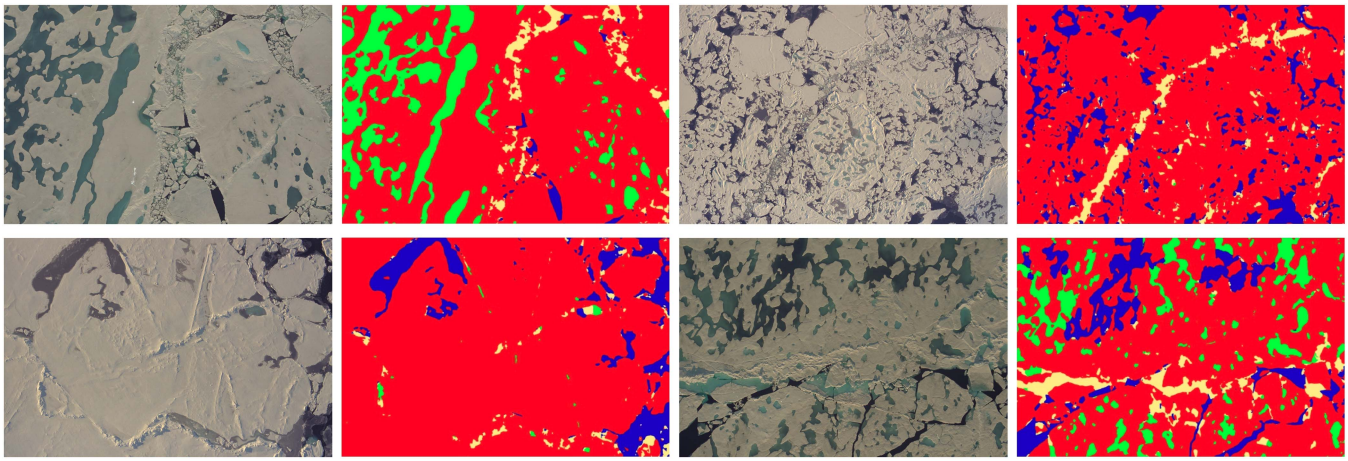


Fig. 7. MeltPondNet segmentation results on our test dataset. The left-hand side images are the aerial images (columns 1 and 3), and the right-hand side images are the segmentation results (columns 2 and 4).

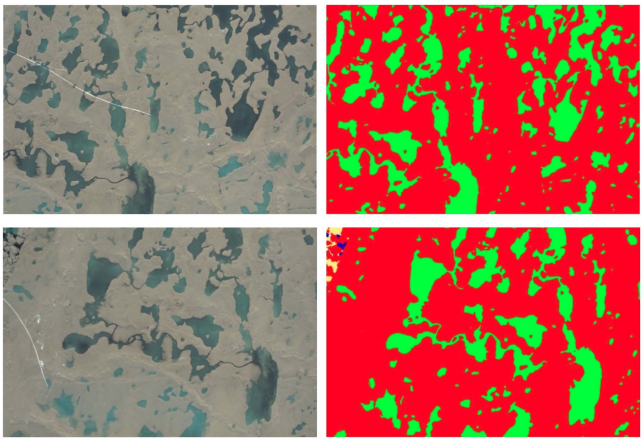


Fig. 8. MeltPondNet segmentation results with image artifacts. The left-hand side images are the aerial images, and the right-hand side images are the segmentation results.

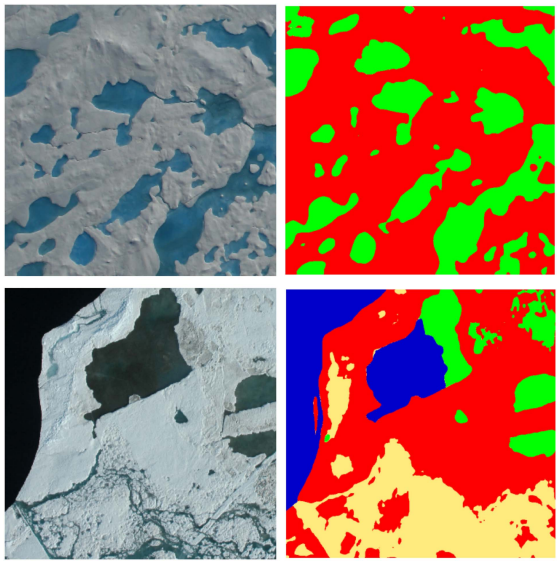


Fig. 9. MeltPondNet segmentation results on high-resolution optical imagery [32]. The left-hand side images are the aerial images, and the right-hand side images are the segmentation results.

the camera may not be working normally, as shown in Fig. 8. Some white lines on the image are considered image artifacts or sensor artifacts. Our MeltPondNet can properly handle those artifacts without losing any accuracy.

5) Robust for Different Sensors: In Fig. 9, the top image is acquired from the optical IceBridge DMS, and the bottom image is acquired from Aerial sRGB [32]. Even though our MeltPondNet is never trained on those data, it can still predict an accurate segmentation result.

V. Conclusion

In this study, we proposed a segmentation algorithm based on a swin transformer U-Net for accurately extracting the boundaries of melt ponds in the surface of Arctic sea ice that operates in high spatial resolution aerial photographs. The model has been trained and assessed using reference data obtained by expert melt pond mapping from aerial photographs taken over sea ice during the HOTRAX in the central Arctic. The mapping has been performed based on albedo differences for four classes of surface: 1) melt pond; 2) ice/snow; 3) submerged ice; and 4) open water. These classes have been used for the model training and application, and we observed the efficiency of separation of melt ponds from other surfaces.

The developed method can be applied not only for melt ponds detection, but their corresponding changes between years that are beneficial for climate studies. The workflow can be adapted for other types of optical data or potentially the data acquired at frequencies in other bands that can extend the algorithm application and provide new insights into processes between ocean and atmosphere. The obtained results discussed in this study are promising and future work could include a comprehensive assessment of the algorithm accuracy in complex climatic transformations. The MeltPondNet architecture would also offer the potential for efficient image analysis of geometrically sophisticated tundra lakes on permafrost [33].

Acknowledgment

Time series photographs were provided by Prof. Donald K. Perovich (personal communication).

APPENDIX A
IMAGE ANALYSIS METHODS: MACHINE LEARNING

Techniques	Class	Goal	Data Source
Maximum likelihood method [34]	Melt ponds open water	1) To evaluate the aerial photographs. 2) To determine the spatial variability of open water, melt pond and snow-covered ice. 3) To determine the area-wise average albedo and ice concentration compare.	Helicopter flights with photographic surveys (CHINARE2010)
Decision tree and random forest based on polarimetric parameters [6]	Open water sea ice melt ponds	1) To develop a novel approach to the retrieval of melt ponds. 2) To derive accurate pond statistics using various polarimetric parameters. 3) To investigate the robustness of the pond statistics obtained from high-resolution multi-polarization SAR data.	The TerraSAR-X dual-polarization data and airborne SAR images
Multilayer perceptron based on the differences of the spectral curves [8]	Open water snow and ice melt ponds	To quantify the surface fractions for melt ponds, open water, and snow and ice for the entire Arctic region for a time period from 2000-2011.	MODIS data
Multilayer neural network and multinomial logistic regression [9]	Melt ponds and ice	To retrieve pan-Arctic binary melt pond classification and melt pond fraction for a time period from 2001-2019.	MODIS data

APPENDIX B
IMAGE ANALYSIS METHODS: IMAGE PROCESSING

Techniques	Class	Goal	Data Source
Otsu's method [35]	Melt ponds	To develop an algorithmic method of mapping a configuration of melt ponds onto a graph of nodes and edges.	Aerial images of Arctic sea ice from the SHEBA and HOTRAX databases
Thresholding [11]	Melt ponds snow cover	To quantify the daily changes in fractional melt-pond coverage over an intensive study area by using a time series of photographic infrared imagery collected from a tethered balloon at an altitude of 300 m.	Photographs acquired during the Seasonal Sea Ice Monitoring and Modelling Site (SIMMS) field experiment of 1995
Combination of different bands [12]	Melt ponds open water	To investigate the possibility to obtain f_{mp} estimates.	ENVISAT WSM images with HH-polarization
Image processing software: ENVI@EX [7]	Melt ponds	1) To map melt ponds using SAR systems. 2) To derive fraction, size, and shape data for melt ponds.	Helicopter-based airborne SAR survey in the northern Chukchi Sea during summer 2011.

REFERENCES

[1] J. C. Stroeve, T. Markus, L. Boisvert, J. Miller, and A. Barrett, "Changes in Arctic melt season and implications for sea ice loss," *Geophysical Res. Lett.*, vol. 41, no. 4, pp. 1216–1225, 2014, doi: 10.1002/2013GL058951.

[2] C. W. Thackeray and A. Hall, "An emergent constraint on future Arctic sea-ice albedo feedback," *Nat. Climate Change*, vol. 9, no. 12, pp. 972–978, Dec. 2019, doi: 10.1038/s41558-019-0619-1.

[3] M. A. Webster, I. G. Rigor, D. K. Perovich, J. A. Richter-Menge, C. M. Polashenski, and B. Light, "Seasonal evolution of melt ponds on Arctic sea ice," *J. Geophysical Res.: Oceans*, vol. 120, no. 9, pp. 5968–5982, 2015, doi: 10.1002/2015JC011030.

[4] D. Flocco, D. Schroeder, D. L. Feltham, and E. C. Hunke, "Impact of melt ponds on Arctic sea ice simulations from 1990 to 2007," *J. Geophysical Res.: Oceans*, vol. 117, no. C09032, 2012, doi: 10.1029/2012JC008195.

[5] S. Kern, A. Rösel, L. T. Pedersen, N. Ivanova, R. Saldo, and R. T. Tonboe, "The impact of melt ponds on summertime microwave brightness temperatures and sea-ice concentrations," *Cryosphere*, vol. 10, no. 5, pp. 2217–2239, 2016.

[6] H. Han et al., "Retrieval of melt ponds on Arctic multiyear sea ice in summer from TerraSAR-X dual-polarization data using machine learning approaches: A case study in the Chukchi sea with mid-incidence angle data," *Remote Sens.*, vol. 8, no. 1, 2016, Art. no. 57. [Online]. Available: https://www.mdpi.com/2072-4292/8/1/57

[7] D.-j. Kim, B. Hwang, K. H. Chung, S. H. Lee, H.-S. Jung, and W. M. Moon, "Melt pond mapping with high-resolution SAR: The first view," *Proc. IEEE*, vol. 101, no. 3, pp. 748–758, Mar. 2013.

[8] A. Rösel, L. Kaleschke, and G. Birnbaum, "Melt ponds on Arctic sea ice determined from modis satellite data using an artificial neural network," *Cryosphere*, vol. 6, no. 2, pp. 431–446, 2012. [Online]. Available: https://tc.copernicus.org/articles/6/431/2012/

[9] S. Lee, J. Stroeve, M. Tsamados, and A. L. Khan, "Machine learning approaches to retrieve pan-Arctic melt ponds from visible satellite imagery," *Remote Sens. Environ.*, vol. 247, 2020, Art. no. 111919.

[10] D. Perovich, T. Grenfell, B. Light, and P. Hobbs, "Seasonal evolution of the albedo of multiyear Arctic sea ice," *J. Geophysical Res.: Oceans*, vol. 107, no. C10, pp. SHE 20-1–SHE 20-13, 2002.

[11] C. Derksen, J. Piwowar, and E. LeDrew, "Sea-ice melt-pond fraction as determined from low level aerial photographs," *Arctic Alpine Res.*, vol. 29, no. 3, pp. 345–351, 1997, doi: 10.1080/00040851.1997.12003254.

[12] M. Mäkynen, S. Kern, A. Rösel, and L. T. Pedersen, "On the estimation of melt pond fraction on the Arctic sea ice with ENVISAT WSM images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7366–7379, Nov. 2014.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[15] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[16] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[19] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

[20] D. K. Perovich et al., "Transpolar observations of the morphological properties of Arctic sea ice," *J. Geophysical Res.: Oceans*, vol. 114, no. C00A04, 2009, doi: 10.1029/2008JC004892.

[21] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Proc. Workshops 26th AAAI Conf. Artif. Intell.*, 2012.

[22] R. Liu, J. Shen, Q. Sun, J. Yang, and S.-C. Cheung, "Cascaded pose regression revisited: Face alignment in videos," in *Proc. IEEE 3rd Int. Conf. Multimedia Big Data*, 2017, pp. 291–298.

[23] B. Adhikari and H. Huttunen, "Iterative bounding box annotation for object detection," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 4040–4046.

[24] T. H. Aspiras, R. Liu, and V. K. Asari, "Active recall networks for multiperspectivity learning through shared latent space optimization," in *Proc. 11th Int. Joint Conf. Comput. Intell.*, 2019, pp. 434–443.

[25] R. Liu and V. K. Asari, "Cascaded semi-supervised annotation tool for image data labeling," *Acta Sci. Comput. Sci.*, vol. 4, pp. 10–15, 2022.

[26] R. Liu, J. Shen, H. Wang, C. Chen, S.-C. Cheung, and V. K. Asari, "Enhanced 3D human pose estimation from videos by using attention-based neural network with dilated convolutions," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1596–1615, 2021.

[27] T. Aspiras, R. Liu, and V. K. Asari, "Convolutional auto-encoder for vehicle detection in aerial imagery (conference presentation)," *Proc. SPIE*, vol. 10995. SPIE, 2019, Art. no. 109950D.

[28] R. Liu, T. H. Aspiras, and V. K. Asari, "Deep neural network based approach for robust aerial surveillance," *Proc. SPIE*, vol. 11735, pp. 57–66, 2021.

[29] R. Liu, T. Aspiras, and V. K. Asari, "Deep neural machine for multimodal information fusion," *Proc. SPIE*, vol. PC12101, 2022, Art. no. PC121010C.

[30] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

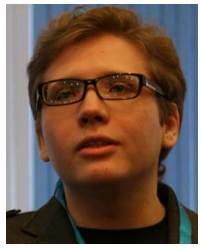
[31] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[32] N. C. Wright and C. M. Polashenski, "Open-source algorithm for detecting sea ice surface features in high-resolution optical imagery," *Cryosphere*, vol. 12, no. 4, pp. 1307–1329, 2018.

[33] I. Sudakov, A. Essa, L. Mander, M. Gong, and T. Kariyawasam, "The geometry of large Tundra lakes observed in historical maps and satellite images," *Remote Sens.*, vol. 9, no. 10, 2017, Art. no. 1072.

[34] L. Li, C. Ke, H. Xie, R. Lei, and A. Tao, "Aerial observations of sea ice and melt ponds near the North Pole during CHINARE2010," *Acta Oceanologica Sinica*, vol. 36, no. 1, pp. 64–72, Jan. 2017, doi: 10.1007/s13131-017-0994-2.

[35] M. Barjatia, T. Tasdizen, B. Song, C. Sampson, and K. M. Golden, "Network modeling of arctic melt ponds," *Cold Regions Sci. Technol.*, vol. 124, pp. 40–53, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165232X15002992



Ivan Sudakow received the master's degree in physics from Ural State University, Ekaterinburg, Russia, in 2008, and the Ph.D. degree in applied mathematics from Novgorod State University, Veliky Novgorod, Russia, in 2012.

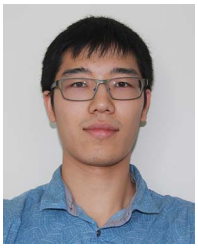
He has been an Assistant Professor with the Department of Physics, University of Dayton, for a long time. He is currently a Lecturer of Applied Mathematics with the School of Mathematics and Statistics, The Open University, Milton Keynes, U.K. He is also a Scholar with the Kavli Institute for Theoretical Physics, Santa Barbara, CA, USA. He specializes in data analysis and mathematical modeling for physical and living systems.

Dr. Sudakow was awarded by German Federal Government the title "Green Talent" in 2013 for "his outstanding research of sea ice and his strong commitment to interdisciplinary interaction between mathematics and climate science".

Vijayan K. Asari (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Indian Institute of Technology Madras, Chennai, India, in 1994.

He is currently a Professor of Electrical and Computer Engineering and the Ohio Research Scholars Endowed Chair in wide area surveillance with the University of Dayton, Dayton, OH, USA, where he is also the Director of the Center of Excellence for Computational Intelligence and Machine Vision (Vision Lab). He holds four U.S. patents and has authored or coauthored more than 700 research articles, including an edited book in wide area surveillance and 116 peer-reviewed journal papers in the areas of image processing, pattern recognition, machine learning, deep learning, and artificial neural networks.

Prof. Asari is an elected Fellow of SPIE and a Co-Organizer of several SPIE and IEEE conferences and workshops. He was the recipient of several teaching, research, advising, and technical leadership awards, including the University of Dayton School of Engineering Vision Award for Excellence in August 2017, the Outstanding Engineers and Scientists Award for Technical Leadership from The Affiliate Societies Council of Dayton in April 2015, and the Sigma Xi George B. Noland Award for Outstanding Research in April 2016.



Ruixu Liu (Member, IEEE) received the B.S. degree in electrical engineering from the Nanjing University of Science and Technology, Nanjing, China, in 2012, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Dayton, Dayton, OH, USA, in 2019 and 2014, respectively.

He is currently a Research Scientist with the University of Dayton. His research interests include computer vision, object detection and segmentation, human pose estimation, and 3-D human pose reconstruction.

Denis Demchev (Member, IEEE) received the Ph.D. degree in oceanography from Saint Petersburg State University, St. Petersburg, Russia, in 2007.

From 2007 to 2019, he was a Leading Programmer with the Center for Ice and Hydrometeorological Information, Arctic and Antarctic Research Institute, and a Researcher with the Nansen International and Remote Sensing Center, St. Petersburg. In 2019, he was a Project Assistant with the Chalmers University of Technology, Gothenburg, Sweden, where he is currently with the Department of Earth, Space, and Environment. From 2020 to 2022, he was a Researcher with the Nansen Environmental and Remote Sensing Center, Bergen, Norway. His research interests include sea ice dynamics retrieval and validation from satellite SAR data, with a focus on localization of ice deformation features, such as leads and ridges.