

River Ice Segmentation with Deep Learning

Abhineet Singh

Hayden Kalke

Mark Loewen

Nilanjan Ray

University of Alberta

asingh1, kalkel, mrloewen, nray1@ualberta.ca

Abstract—This paper deals with the problem of computing surface ice concentration for two different types of ice from digital images of river surface. It presents the results of attempting to solve this problem using several state of the art semantic segmentation methods based on deep convolutional neural networks (CNNs). This task presents two main challenges - very limited availability of labeled training data and presence of noisy labels due to the great difficulty of visually distinguishing between the two types of ice, even for human experts. The results are used to analyze the extent to which some of the best deep learning methods currently in existence can handle these challenges. The code and data used in the experiments are made publicly available to facilitate further work in this domain.

I. INTRODUCTION

The study of surface ice concentration and variation over time and place is crucial for understanding the process of river ice formation. The computation of temporal and spatial ice distributions can help to validate models of this process. The additional ability to distinguish frazil ice from the sediment-carrying anchor ice can also increase the estimation accuracy of the sediment transportation capacity of the river. Towards this end, a large amount of video data has been captured using UAVs and bridge mounted game cameras from two Alberta rivers during the winters of 2016 and 2017. The objective of this work is to analyze this data and perform dense pixel wise segmentation on these images to automatically compute the concentrations of the two types of ice.

The main challenge in this task is the lack of labeled data since it is extremely difficult and time consuming to manually segment images into the three categories, owing to the arbitrary shapes that the ice pans can assume. As a result, there are currently only 50 labeled images (Fig. 1) to accompany 564 unlabeled test images and over 100 minutes of unlabeled high resolution (4K) videos. These labeled images along with 205 additional images with only ice-water labeling had previously been used to train an SVM [22], [21], [20] to perform segmentation that provided water-ice classification accuracies ranging from 80.1% - 93.5% and surface ice concentration errors of 0.7% - 3.3%. Though these methods were fairly successful at separating ice from water, they had difficulty in distinguishing between frazil and anchor ice pans, especially in cases where they are not physically separated and are hard to differentiate, even for human eyes. This paper is mainly concerned with handling these more difficult cases.

To address the limitations of SVM-based ice classification, this work uses recent semantic segmentation methods based on deep CNNs. Since CNNs need large amounts of training data to work well, several data augmentation techniques (Sec. III-C) have been used to generate enough training images. Detailed

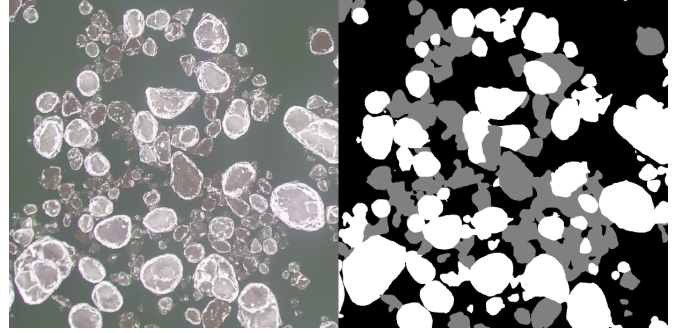


Fig. 1: A sample training image with corresponding label where white, gray and black pixels respectively denote frazil ice, anchor ice, and water.

ablation studies have also been performed (Sec. IV-B2, IV-B3) to evaluate the extent to which few and partially labeled training images can be used to obtain results that are conducive to generating further training data with minimum human effort and thus setup a bootstrapping process.

II. BACKGROUND

Surface ice is formed on rivers in cold regions like Canada when the air temperature remains below freezing for extended periods of time. It begins with the formation of frazil ice crystals and flocs as water columns become supercooled [11]. Being naturally adhesive [23], these ice crystals sinter together to form larger flocs whose increasing buoyancy brings them to the surface as slush. As the moving slush is in turn subjected to freezing temperatures, it forms solid frazil pans and rafts which can become circular with upturned edges on colliding with each other. Examples of these formations can be seen in Fig. 1.

Anchor ice is formed when frazil flocs come in contact with the river bed or other solid surface and freeze there to become immobile. Such accumulations grow in multiple stages [25] to reach vertically towards the surface. This mechanism increases the drag on the top of the accumulation which, along with internal buoyancy of the formation and the weakening effect of incoming solar radiation, can eventually overcome the bond between the formation and the substrate, causing it to break apart and float to the surface as anchor ice pans. These pans carry entrapped sediments and other materials from the river bed and transport them downstream.

An improved understanding of the importance of such transportation to the overall sediment budget of the river would be useful for developing and validating models of river processes. Thus, we require accurate estimates for the concentrations

of sediment-carrying anchor ice in the river for which field measurements are unavailable. Since there is far too much data for manual analysis to be practical, one of the objectives of this work is to estimate this concentration from digital images and videos of the river surface in an automated or semi-automated manner using deep learning. Distinguishing ice from water is relatively straightforward and has been accomplished fairly successfully using simple techniques like thresholding [2] as well as classic machine learning methods like SVM with handcrafted features [22], [21], [20]. The main goal of this work is therefore to be able to distinguish between frazil and anchor ice pans with high accuracy using state of the art deep learning techniques.

River ice images bear a significant resemblance to microscopic images of cells in the bloodstream which initially suggested the use of existing cell classification networks from medical imaging. There are several promising studies employing a range of architectures including ConvNet [40], LeNet [34], [30], Resnet [26] and Inception [27] that might have provided the base networks for such an approach. However, a more detailed examination of these studies revealed that medical imaging tasks are mainly concerned with the detection and localization of specific kinds of cells rather than performing pixel wise segmentation that is necessary to estimate ice concentration.

Further, unsupervised and semi-supervised video segmentation techniques were considered to better utilize the large amount of unlabeled but high-quality video data that has been collected. Most of these methods use optical flow for performing motion segmentation [13], though some appearance based [4] and hybrid [15], [19] methods have also been proposed. An unsupervised bootstrapping approach has been proposed in [31] where motion segmented images are used as training data to learn an implicit representation of this object under the assumption that all moving pixels belong to the same foreground object. This learnt model is then used to refine the motion segmentation and the improved results are in turn used to bootstrap further refinements.

Unfortunately, two underlying assumptions in [31], and motion segmentation, in general, render such methods unsuitable for the current task. First, they assume that there is a single moving foreground object whereas the objective here is to distinguish between two different types of moving ice, both of which are foreground objects. Second, they assume a static background while the river, which makes up the background here, is itself moving. Preliminary attempts to perform optical flow-based motion segmentation on the river ice videos confirmed its unsuitability for this work. There is a method [37] for performing simultaneous optical flow estimation and segmentation which might be able to address these limitations to some extent. However, it had only a Matlab implementation available that was far too slow for our purpose so its further exploration was deferred to future work.

Finally, it seems that very limited existing work has been done on the application of deep learning for surface ice analysis as the only one that was found [8] uses microwave sensor data instead of images.

TABLE I: Trainable parameter counts for the four models

Model	Deeplab	UNet	SegNet	DenseNet
Parameters	41,050,787	12,284,019	11,546,739	90,948

III. METHODOLOGY

A. Data Collection and Labeling

Digital images and videos of surface ice conditions were collected from two Alberta rivers - North Saskatchewan River and Peace River - in the 2016-2017 winter seasons. Images from North Saskatchewan River were collected using both Reconyx PC800 Hyperfire Professional game cameras mounted on two bridges in Edmonton as well as a Blade Chroma UAV equipped with a CGO3 4K camera at the Genesee boat launch. Data for the Peace River was collected using only the UAV at the Dunvegan Bridge boat launch and Shaftesbury Ferry crossing. The game camera captured 3.1 megapixels resolution still images at one-minute frequency while the UAV camera captured 4K videos (8.3 megapixels) of up to 10 minutes duration.

Large 3840×2160 UAV images were cropped into several 1280×1080 images to make labeling more convenient while the smaller 2048×1536 game camera images were only cropped to remove text information added by the camera software. More than 200 of these images were labeled for binary ice-water classification but only 50 of these were labeled into 3 classes to distinguish between the two types of ice and the water. Only the latter images were used for training in this work. More details of the data collection and labeling process along with images of the camera setups are available in [21, sec. 4.1].

B. Image Segmentation

Since neither cell classification nor video segmentation methods seemed promising, it was decided to rely only on supervised image segmentation. After extensive research through several excellent resources for these methods [29], [12], four of the most widely cited and best performing methods with publicly available implementations were selected. Descriptions of these methods that follow have been kept brief and high-level because of the empirical nature of this work and the target audience.

The first of these models is UNet [32] from the medical imaging community. It was introduced for neuronal structure segmentation in electron microscopic images and won the ISBI challenge 2015. As the name suggests, UNet combines a contracting part with a symmetric expanding part to yield a U-shaped architecture that can both utilize contextual information and achieve good localization owing to the two parts respectively. It was shown to be trainable with relatively few training samples while relying heavily on patch based augmentation which seemed to make it an ideal fit for this study.

The second network is called SegNet [3] and was introduced for segmenting natural images of outdoor and indoor scenes for scene understanding application. It uses a 13-layer VGG net [35] as its backbone and features a somewhat similar architecture as UNet. The contracting and expanding parts

are here termed as encoder and decoder, respectively and the upsampling units in the latter are not trainable, instead sharing weights with the corresponding max-pooling layers in the former. Keras [10] implementations were used for both UNet and SegNet, available as part of the same repository [16].¹

The third method is called Deeplab [5] and is one of the best performing methods in the Tensorflow research models repository [6]. It uses convolutions with upsampled filters - the so called atrous convolutions [38] - both to achieve better control over the feature response resolution and to incorporate larger context without increasing the computational cost. It also uses pyramidal max pooling to achieve scale-invariance and combines its last layer output with a fully connected conditional random field layer to improve localization accuracy while maintaining spatial invariance. This work uses a more recent version called Deeplabv3+ [7] which adds a decoder module to produce sharper object boundaries and uses the powerful Xception backbone architecture [9] for further performance improvements.

The fourth method is based on the DenseNet architecture [17]. To the best of our knowledge, this architecture has not yet been applied for segmentation but is included here due to its desirable property of providing state of the art performance with a much smaller network size. The basic idea of DenseNet is to connect each hidden layer of the network to all subsequent layers so that the feature maps output by each layer are used as input in all subsequent layers. This provides for better multi-resolution feature propagation and reuse while drastically reducing the total number of parameters and mitigating the vanishing gradient problem. The architecture used in this work had 9 such layers though experiments were done with the more layers up to 21 (Sec. III-C). As shown in Table I, DenseNet has by far the fewest parameters of the four models, being over two orders of magnitude smaller than the next smallest one.

C. Data Augmentation and Training

A simple sliding window approach was used to extract a large set of sub-images or patches from each training image. The window was moved by a random stride between 10% to 40% of the patch size K . This process was repeated after applying random rotations to the entire image between 15 to 345 degrees divided into four bands of equal width to allow for multiple rotations for each image. Finally, each patch was also subjected to horizontal and vertical flipping to generate two additional patches. All resultant patches were combined together to create the dataset for each K . For testing a model, patches of size K were extracted from the test image using a stride of K , segmentation was performed on each patch and the results were stitched back to get the final result.

All models were trained and tested using patch sizes $K \in \{256, 384, 512, 640, 800, 1000\}$. DenseNet turned out to perform best with $K = 800$ while all other models did so

with $K = 640$. All results in Sec. IV were therefore obtained using these patch sizes. The 50 labeled images were divided into two sets of 32 and 18 for generating the training and testing/validation images, respectively. Results on some of the unlabeled videos (Sec. IV-C2) were also generated using models trained on all 50 images.

UNet and SegNet were both trained for 1000 epochs and the training and validation accuracies were evaluated after each. The trained model used for testing was the one with either the maximum validation accuracy or the maximum mean accuracy depending on how well the training and validation accuracies were matched in the two cases. Deeplab was trained for between 100,000 and 200,000 steps. Batch size of 10 was used for $K = 256$ and 2 for $K \in \{640, 800, 1000\}$ with the latter chosen due to memory limitations. $K = 384$ was tested with batch sizes 6 and 8 while $K = 512$ was tested with 6 and 2. Most tests were conducted using the default stride of 16 with corresponding atrous rates of [6, 12, 18] though one model with $K = 256$ was also trained using Stride 8 with atrous rates of [12, 24, 36].

DenseNet training was a bit more complicated. Simply using all the pixels for training caused the network to rapidly converge to a model that labeled all pixels with the class that had the most training pixels - water in most cases. To get meaningful results, the number of pixels belonging to each of the classes had to be balanced. Therefore 10,000 random pixels belonging to each class were selected in each epoch, with different sets of pixels selected each time, and only these were used for computing the loss. Training images with less than 10,000 pixels in any class were discarded. The number of epochs were between 1000 – 1600 for all K . In all cases, the performance metrics in Sec. IV-A were computed on the validation set every 10 epochs and training was stopped when these became high enough or remained unchanged for over 100 epochs.

D. Ablation Experiments

One of the principal difficulties in training deep models for performing segmentation is the lack of sufficient labeled data due to the tedious and time-consuming manual segmentation of images. This problem is exacerbated in the current task because of the difficulty in distinguishing between the two types of ice that exhibit both high intraclass variation and significant appearance overlap in addition to arbitrary and difficult to delineate shapes. As a result, a highly desirable attribute of a practically applicable model would be its ability to learn from a few images including partially labeled ones.

Two different types of ablation experiments were performed in order to explore the suitability of the tested models in this regard. The first one was to train the models using different subsets of the training set. The second one was to consider the labels from only a small subset of pixels in each image to simulate the scenario of partially labeled training data. Note that the input image itself was left unchanged so that the models did have access to all the pixels but the loss function minimized during training was computed using only the labels from the selected pixels.

¹This repository also includes two variants of the FCN architecture [28], [33] that were also tested but did not perform as well as the other two and are thus excluded here.

SegNet exhibited similar performance patterns as UNet in the ablation tests, while being slightly worse on average, probably because they share the same base network. SegNet results have thus been excluded in Sec. IV-B2 and IV-B3 for the sake of brevity. All the code and data used for training, augmentation and ablation tests is **made publicly available** [36][1] along with detailed instructions for using it to facilitate further work in this domain and easy replication of results.

IV. RESULTS

A. Evaluation Metrics

Following evaluation metrics are typically used in image segmentation [33], [24]:

1) Pixel accuracy:

$$\text{pix_acc} = \frac{\sum_i n_{ii}}{\sum_i t_i} \quad (1)$$

2) Mean accuracy:

$$\text{mean_acc} = \frac{1}{n} \sum_i \frac{n_{ii}}{t_i} \quad (2)$$

3) Mean IOU:

$$\text{mean_iou} = \frac{1}{n} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \quad (3)$$

4) Frequency Weighted IOU:

$$\text{fw_iou} = \sum_k (t_k)^{-1} \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \quad (4)$$

where n is the number of classes, n_{ij} is the number of pixels of class i predicted to belong to class j and t_i is the total number of pixels of class i in the ground truth. Note that accuracy measures only the rate of true positives while IOU also accounts for false positives. Hence, **accuracy and IOU are respectively equivalent to recall and precision** that are more widely used in pattern recognition and are referenced thus in the remainder of this paper. Note also that pix_acc is a frequency weighted version of mean_acc and is thus referred to as **frequency weighted (fw) recall** in Table III.

Using these metrics to measure the combined segmentation performance over all three classes can lead to biased results when the number of image pixels is not evenly distributed between the classes. This is particularly so in the current work whose main objective is to distinguish between the two types of ice. However, as shown in Table II, more than half the pixels in both sets of test images are of water while anchor ice, which is the most difficult to segment, covers only about 17% of the pixels. Therefore, results in the next section are mostly restricted to class specific versions of these metrics.

TABLE II: Class frequencies in the two test sets

test images	water	anchor ice	frazil ice
18	0.557	0.175	0.268
46	0.610	0.166	0.224

In addition to these segmentation metrics, the estimated ice concentration accuracy has also been used. This is computed

through a three-step process. First, column-wise ice concentration is obtained for each frame by computing the percentage of all pixels in each column that are classified as ice (combined, anchor or frazil). These values are stacked together for all columns from left to right to form a vector of the same size as the frame width. Next, the mean absolute error (MAE) is computed between the ice concentrations vectors produced by the ground truth and a model to obtain the overall accuracy for that frame. Finally, the median of MAE values over the entire test set is taken as the final metric. Median has been preferred over mean as being significantly more robust to outliers.

The above metrics are only applicable to labeled images so that unlabeled videos can only be evaluated qualitatively but the conclusions thus obtained are usually somewhat subjective. In order to ameliorate this, an unsupervised metric named **mean ice concentrations difference** has been proposed to measure the consistency of segmentation results between consecutive video frames as a proxy for its accuracy. This metric is computed as the mean absolute difference between the ice concentration vectors of each pair of consecutive frames in the video. Segmentation consistency over the entire video is summarized by taking the average of these differences over all pairs of consecutive frames. The intuition behind this metric is that, since the ice and/or river are moving slowly and video FPS is fairly high, the ice concentration changes very gradually and its difference between consecutive frames remains small. A model that fails to generalize well to the videos would give inconsistent results in corresponding patches from nearby frames that would therefore result in a high mean concentration difference. Experiments were also done using the direct pixel-wise difference between the segmentation masks themselves, both with and without incorporating motion estimation by optical flow [39], [14], [18], but the results were less consistent so these were excluded here.

B. Quantitative Results

1) *Overview:* As shown in Table III, all of the deep models provide significant improvement over SVM for all cases except a couple instances of frazil ice. This is most notable for anchor ice where an increase of 12 – 20% in recall and 6 – 20% in precision is achieved in absolute terms, with the respective relative increases being 19 – 34% and 13 – 44%. It is noteworthy that the two best models Deeplab and UNet provide greater performance improvement, in both absolute and relative terms, with respect to precision than recall over all 4 categories. This is particularly impressive since high precision is usually harder to achieve than recall, as testified by its lower values across all models and categories.

Further, we see that DenseNet and SegNet fall slightly behind SVM on frazil ice, especially with respect to recall, even though they have the two highest recalls on anchor ice. This trend of an inverse relationship between anchor and frazil ice recall was consistently observed in the ablation tests too (Sec. IV-B2). It seems that learning to better distinguish anchor from frazil ice often comes at the cost of either a decrease in the capability to recognize frazil ice itself or an overcorrection which causes some of the more ambiguous cases of frazil to

TABLE III: Segmentation recall and precision for SVM and all deep models trained and tested on the 32 and 18 image sets respectively. The fw in the last category stands for frequency weighted and the corresponding recall and precision metrics refer to pix_acc (Eq. 1) and fw_iou (Eq. 4) as detailed in Sec. IV-A. Relative increase over SVM is computed as $(\text{model_value} - \text{svm_value}) / \text{svm_value} \times 100$

Model	anchor ice		frazil ice		ice+water		ice+water (fw)	
	Metric Value (%)	Relative increase over SVM	Metric Value (%)	Relative increase over SVM	Metric Value (%)	Relative increase over SVM	Metric Value (%)	Relative increase over SVM
Recall								
svm	61.54	-	75.41	-	78.12	-	84.93	-
deeplab	74.46	21.00	87.51	16.05	86.38	10.57	90.87	7.00
unet	73.75	19.85	84.27	11.75	85.13	8.97	88.69	4.42
densenet	76.96	25.06	71.06	-5.77	81.42	4.22	85.02	0.11
segnet	82.31	33.75	68.99	-8.51	83.06	6.32	85.90	1.14
Precision								
svm	43.32	-	63.07	-	65.84	-	76.84	-
deeplab	62.39	44.03	77.14	22.32	77.25	17.33	84.32	9.72
unet	54.89	26.72	71.17	12.84	73.19	11.17	81.73	6.36
densenet	48.98	13.07	60.97	-3.32	67.69	2.82	77.49	0.84
segnet	52.80	21.90	62.60	-0.73	69.60	5.72	78.46	2.10

be misclassified as anchor ice. It is likely that the loss function can be minimized equally well by over-fitting either to frazil or to anchor ice, thus leading to two stable training states.

Comparing between the deep models themselves, Deeplab turns out to be the best overall followed closely by UNet. It is interesting to note that, while DenseNet and SegNet provide better recall with anchor ice, their corresponding precision is lower. Since recall does not penalize false positives while precision does, this probably indicates that DenseNet and SegNet misclassify frazil as anchor ice more often than Deeplab and UNet, which is consistent with the inverse relationship hypothesis. Further, it can be seen that the performance difference between deep models and SVM decreases when all three classes are considered and even more so when the averaging is frequency weighted. As mentioned before, these are the cases where high segmentation accuracy of water starts to dominate. Finally, the greater difficulty of recognizing anchor ice over frazil ice is confirmed by its significantly lower recall and precision in almost all cases.

Table IV shows the ice concentration estimation accuracy of all the models in terms of median MAE. It should be noted that recall and precision are better indicators of raw segmentation performance since MAE suffers from an averaging effect where false positives can cancel out false negatives to provide an overall concentration value that happens to be closer to the ground truth. As a result, MAE does not always provide a true indication of the actual recognition ability of the model. As an example, SegNet has slightly better MAE than UNet on frazil ice even though it has significantly lower (9 – 16%) precision and recall. Nevertheless, except for the single case of DenseNet with frazil ice, the deep models are consistently better than SVM here too, especially on anchor ice. As with the

TABLE IV: Median MAE for anchor and frazil ice over the 18 test images. Relative decrease over SVM is computed as $(\text{svm_mae} - \text{model_mae}) / \text{svm_mae} \times 100$

Model	anchor ice		frazil ice	
	Median MAE (%)	Relative decrease over SVM	Median MAE (%)	Relative decrease over SVM
svm	8.37	-	7.18	-
deeplab	4.71	43.80	4.52	37.01
unet	6.51	22.29	6.80	5.22
densenet	7.24	13.59	7.20	-0.30
segnet	6.48	22.61	6.64	7.51

segmentation metrics, Deeplab is the best model and provides around 3% absolute and 40% relative improvement over SVM.

2) *Ablation study with training images:* For this study, models were trained using 4, 8, 16, 24 and 32 images and each one was tested using the same 18-image test set. Results for both anchor and frazil ice are given in Fig. 2. Contrary to expectation, a distinct pattern of improvement with more images is not shown by most of the models. There is a slight improvement in anchor ice performance but it seems too weak to clearly demonstrate model improvement with increase in training images. A more likely conclusion is that the test set is just too similar to the training set and does not contain enough challenging variation to allow the extra information from more training images to be reflected in the performance numbers. This is lent some credence by the fact that the 50 labeled images were specifically chosen for their ease of labeling

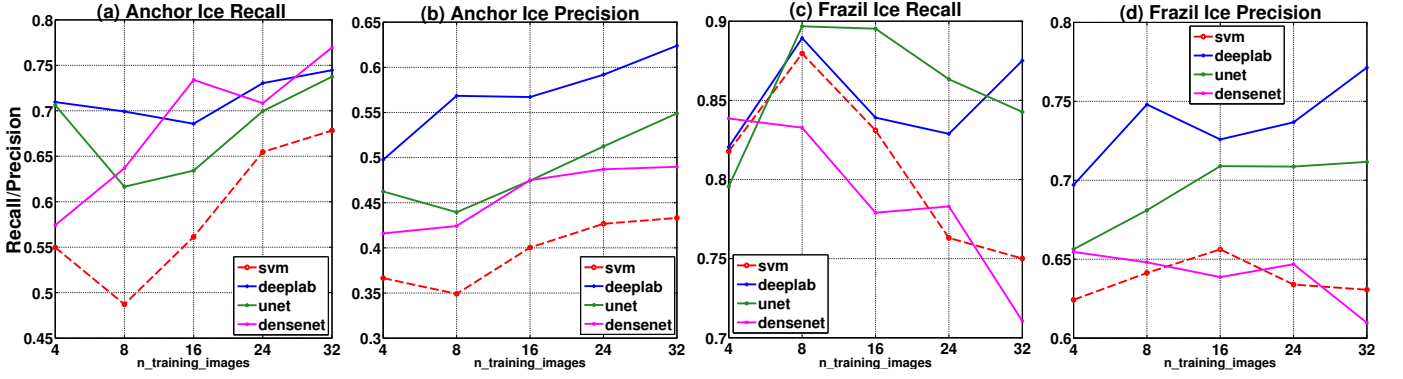


Fig. 2: Results of ablation tests with training images for (a-b) anchor ice and (c-d) frazil ice. Note the variable Y-axis limits.

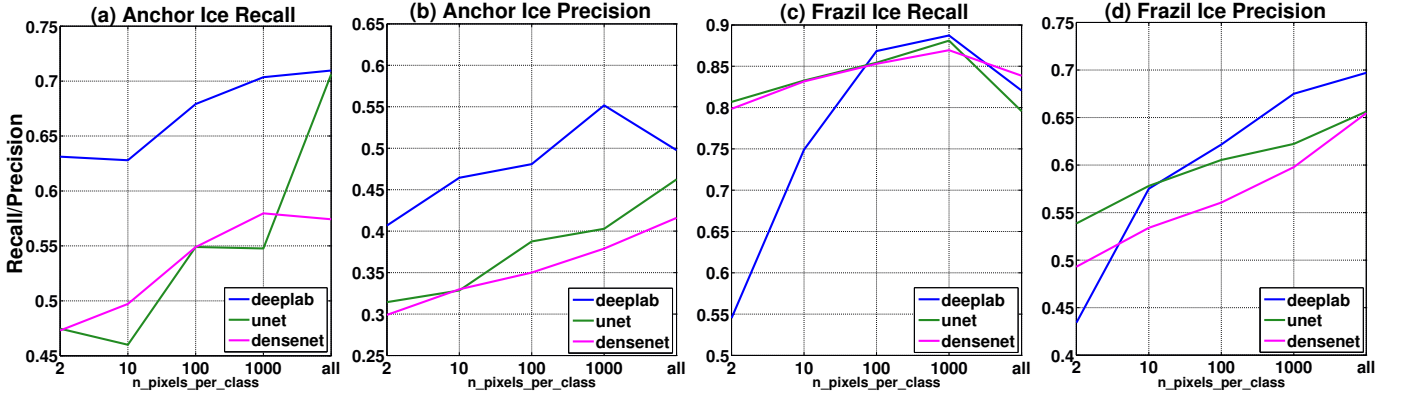


Fig. 3: Results of ablation tests with selective pixels for (a-b) anchor ice and (c-d) frazil ice. Note the variable Y-axis limits.

owing to the highly tedious and time-consuming nature of performing pixel-wise segmentations in addition to the high degree of ambiguity and subjectivity in classifying ice in the more difficult cases. Combined with the fact that they were labeled by the same person, it would not be unusual for them to be similar, both in terms of content and level of challenge.

The inverse relation between anchor and frazil ice recall that was observed in the previous section is apparent here too. For instance, the plot lines for anchor and frazil ice (Fig. 2 (a),(c)) are virtually reflections of each other for all models including SVM except perhaps for Deeplab. Deeplab itself exhibits nearly constant recall though with a clearer upward trend in precision. It is also overall the best model, as in the previous section. SVM shows the strongest improvement in anchor ice recall, along with the corresponding decline in frazil ice while DenseNet does so among the deep models. Also, Fig. 2 (a) and (b) illustrate the superiority of deep models over SVM for anchor ice recognition more clearly than Table. III. Deeplab and UNet largely maintain an appreciable superiority over SVM for frazil ice too (Fig. 2 (c), (d)), though the overall improvement there is less distinct. Finally, DenseNet does seem to be the worst performing deep model, especially for frazil ice, but its competitiveness is still noteworthy considering that it has over two orders of magnitude fewer parameters than the other models (Table I).

3) *Ablation study with selective pixels*: This study was performed by training models using 2, 10, 100 and 1000 pixels per class selected randomly from each $K \times K$ training patch.

The training set was generated from only 4 training images and not subjected to augmentation. Also, $K = 640$ was used for all models including DenseNet to ensure that the number of training pixels remained identical for all of them. Further, in an attempt to counteract the limited challenges available in the 18-image test set, these models were tested on all of the remaining 46 labeled images. Finally, SVM was not included here because its super-pixel based method [22] does not lend itself well to training using randomly selected pixels.

Results are given in Fig. 3. It turns out that selective pixel training has surprisingly little impact on quantitative performance except perhaps in the case of UNet with anchor ice and DeepLab with frazil ice. Though there is a more strongly marked upward trend in performance compared to training images (Fig. 2), it is not as significant as would be expected. The case of 2 pixels per class is particularly remarkable. When combined with the fact that the unaugmented training set contained only 46 patch images, this training was done using only 92 pixels per class or 276 pixels in all. This might be another indicator of the limited challenges available in the test set. This is further confirmed by the qualitative results on videos (Sec. IV-C2) that show a much more strongly marked difference than would be inferred by these plots.

C. Qualitative Results

1) *Images*: Fig. 5 shows the results of applying the optimal configurations of the four models (as used in Sec. IV-B1) to

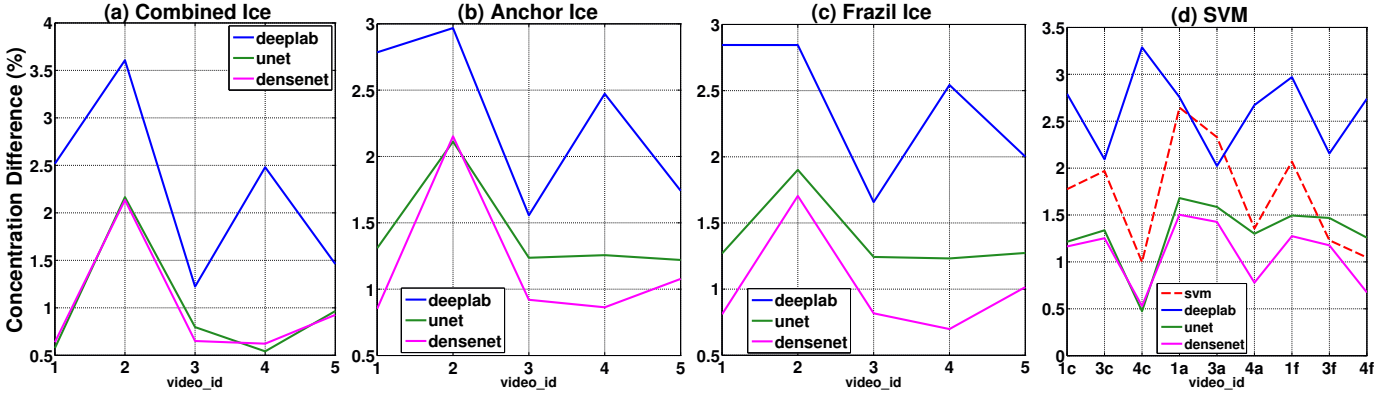


Fig. 4: Mean ice concentration difference between consecutive frames for (a) both types of ice combined, (b) anchor ice, (c) frazil ice and (d) all three for SVM sequences. Details of videos corresponding to the IDs are in the supplementary and suffixes c, a and f in (d) respectively refer to combined, anchor and frazil ice.

segment several images from the unlabeled test set. Additional results are in the supplementary. Several interesting observations can be made. First, both UNet and SegNet misclassify water as frazil ice in several cases where water covers most of the image, e.g. in image 3. DenseNet too seems to be susceptible to this issue, albeit to a much lesser extent, though a careful examination of its video results (Table 4) shows this problem to be more prevalent than the images alone indicate.

Second, Deeplab results show the largest degree of discontinuity between adjacent patches due to its tendency to occasionally produce completely meaningless segmentations on some individual patches. Image 5 is an example.

Third, consistent with the quantitative results of the previous section, DenseNet is overall the best performing model, even though its results are slightly more fragmented than the others. This is particularly noticeable in the more difficult cases of distinguishing between frazil and anchor ice when they both form part of the same ice pan. Images 1 and 7 are examples.

Qualitative results on labeled test image are available in the accompanying data [1] as well as in Google Photos albums whose categorized links are given in the supplementary for convenience.

2) *Videos*: All the deep models were evaluated on 1 to 2 minutes sequences from 5 videos captured on 3 different days and containing wide variations in the scale and form of ice pans. SVM took 5 minutes to process each frame so could only be evaluated on 30 seconds of video 1 and 10 seconds each of videos 3 and 4. Also, selective pixel models were only evaluated on videos 1 and 3. Results for all videos are available in the accompanying data [1]. The supplementary provides details of the tested sequences along with categorized links for some of the results.

The most noticeable point in these results is that Deeplab is susceptible to completely misclassifying individual randomly distributed patches which can lead to strong discontinuities when these patches are stitched together to create complete frames and the corresponding video. This is quantitatively confirmed by Fig. 4 that shows the mean ice concentration difference between consecutive frames in all of the videos for each of the two types of ice as well as both combined. It can

be seen there that Deeplab has significantly higher values than both UNet and DenseNet, being more than 3 times higher in several cases, while DenseNet almost always has the smallest difference, thus indicating the most consistent segmentation results. Apart from confirming the limited challenges available in the labeled test sets, this inversion of relative performance between the 3 models as compared to the quantitative results in Sec. 2, shows that DenseNet is able to generalize to new scenarios much better than Deeplab which has a tendency to overfit to the training images while UNet provides a good balance between generalization and overfitting. Fig. 4 (d) shows the concentration differences over the 3 sequences on which SVM was also tested. SVM classifies pixels in groups of super pixels so it doesn't suffer from the issue of misclassifying entire patches that the deep models are susceptible to but, even so, its generalization ability is poor enough to give significantly higher concentration differences as compared to both DenseNet and UNet, though being slightly better than Deeplab. Moreover, this group classification technique has the disadvantage of giving up blocky appearance to its segmentation masks, whose boundaries are often too coarse to correspond well with the actual ice pans.

Examination of the ablation test videos gives another important result that was mentioned in Sec. IV-B3 - selective pixel training has significantly greater impact in practice than indicated by Fig. 3. The segmentation masks seem to become more grainy and sparse as the number of pixels is decreased and there is a very noticeable difference between using 2 and 1000 pixels. Similarly, a greater difference is apparent between the results produced by models trained on 4 and 32 images than suggested by Fig. 2.

V. CONCLUSIONS AND FUTURE WORK

This paper presented the results of using four state of the art deep CNNs for segmenting images and videos of river surface into water and two types of ice. Three of these - UNet, SegNet and Deeplab - are previously published and well studied methods while the fourth one - DenseNet - is a new method, though based on an existing architecture. All of the models provided considerable improvements over previous attempts

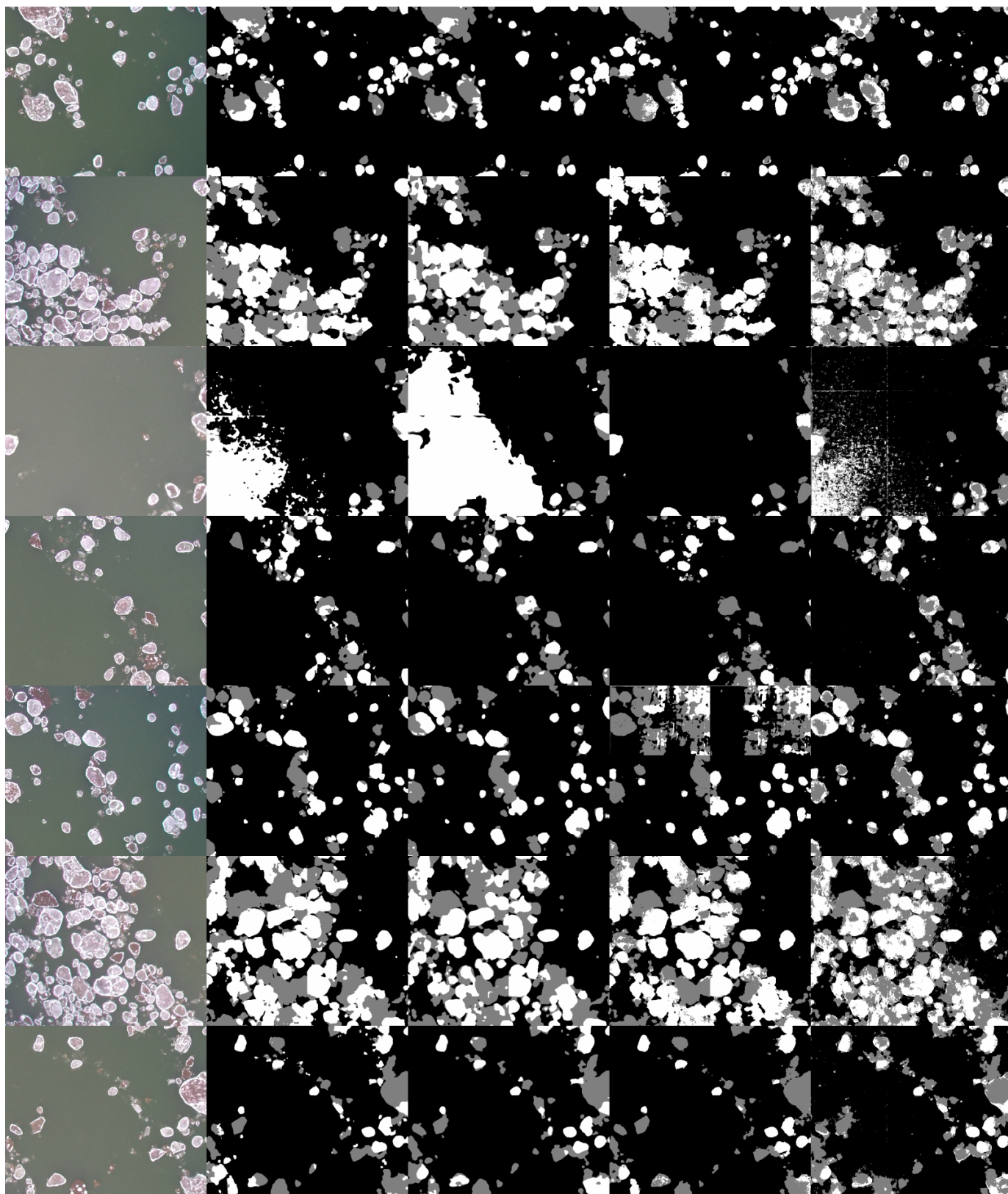


Fig. 5: Results of testing the deep models on unlabeled images: left to right: raw image, UNet, SegNet, Deeplab, DenseNet

using SVM. These were particularly significant for the most challenging case of anchor ice where around 20% increase in both recall and precision were obtained by the four models combined while the single best model - Deeplab - provided respective improvements of 13–19% in absolute and 21–44% in relative terms. Frazil ice performance was slightly less impressive but still surpassed SVM by 12–14% in absolute and 16–22% in relative terms. Significant improvements were obtained in ice concentration estimation accuracy too, with Deeplab providing around 3% absolute and 40% relative decrease in MAE over SVM for both types of ice.

Among the four models, Deeplab gave the best quantitative performance on the labeled test set - 5–10% improvement in precision with similar recall for anchor ice and 3–15% better recall and precision for frazil ice - but showed poor generalization ability by giving the worst qualitative results on the unlabeled images and videos - up to 7-fold mean concentration difference on the videos. DenseNet, on the other hand, gave poor quantitative results but demonstrated excellent generalization ability on the unlabeled data. UNet provided a good balance between the two and might be taken to be the single best model tested here, if such a one needs to be chosen. Finally, this paper also demonstrated reasonable success in handling the lack of labeled images using data augmentation. Further improvements in this direction would be obtained by using the augmented dataset as the starting point for a semi-automated boot-strapping process where segmentation results on unlabeled images would be corrected manually to yield more labeled data to be used for training better models.

REFERENCES

- [1] Alberta River Ice Segmentation Dataset. <http://dx.doi.org/10.21227/ebax-1h44>, 2019. 4, 7
- [2] S. Ansari, C. Rennie, O. Seidou, J. Malenchak, and S. Zare. Automated monitoring of river ice processes using shore-based imagery. *Cold Regions Science and Technology*, 142:1–16, 2017. 2
- [3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017. 2
- [4] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2
- [5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 3
- [6] L.-C. Chen, Y. Zhu, and G. Papandreou. DeepLab: Deep Labelling for Semantic Image Segmentation. Github, 2019. <https://github.com/tensorflow/models/tree/master/research/deeplab>. 3
- [7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv:1802.02611*, 2018. 3
- [8] J. Chi and H.-c. Kim. Prediction of arctic sea ice concentration using a fully data driven deep neural network. *Remote Sensing*, 9(12), 2017. 2
- [9] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1800–1807, 2017. 3
- [10] F. Chollet et al. Keras. <https://keras.io>, 2015. 3
- [11] S. F. Daly. Report on frazil ice. Technical report, International Association for Hydraulic Research Working Group on Thermal Regimes, Hanover, New Hampshire, August 1994. 1
- [12] Dr.Tang. Semantic-Segmentation. Github, 2017. https://github.com/tangzhenyu/SemanticSegmentation_DL. 2
- [13] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In *BMVC*, 2014. 2
- [14] G. Farneback. Two-frame motion estimation based on polynomial expansion. In J. Bigun and T. Gustavsson, editors, *Image Analysis*, pages 363–370. Springer Berlin Heidelberg, 2003. 4
- [15] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *Computer Vision and Pattern Recognition (CVPR 2010)*, 2010. 2
- [16] D. Gupta. Image Segmentation Keras : Implementation of Segnet, FCN, UNet and other models in Keras. Github, 2017. <https://github.com/divamgupta/image-segmentation-keras>. 3
- [17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269. IEEE Computer Society, 2017. 3
- [18] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 4
- [19] S. Jain, B. Xiong, and K. Grauman. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. *CVPR*, 2017. 2
- [20] H. Kalke. Sediment Transport by Released Anchor Ice: Field Measurements and Digital Image Processing . Master’s thesis, University of Alberta, 2017. 1, 2
- [21] H. Kalke and M. Loewen. Support Vector Machine Learning Applied to Digital Images of River Ice Conditions. Submitted for review to Cold Regions Science and Technology, September 2017. 1, 2
- [22] H. Kalke and M. Loewen. Predicting Surface Ice Concentration using Machine Learning . In *19th Workshop on the Hydraulics of Ice Covered Rivers*, Whitehorse, Yukon, Canada., July 9-12 2017. CGU HS Committee on River Ice Processes and the Environment. 1, 2, 6
- [23] E. Kempema and R. Ettema. Anchor ice rafting: Observations from the laramie river. *River Research and Applications*, 27:1126–1135, 11 2011. 1
- [24] M. Ker?ner. Image Segmentation Evaluation. Github, 2017. https://github.com/martinkersner/py_img_seg_eval. 4
- [25] D. J. Kerr, H. T. Shen, and S. F. Daly. Evolution and hydraulic resistance of anchor ice on gravel bed. *Cold Regions Science and Technology*, 35(2):101–114, 2002. 1
- [26] H. Lei, T. Han, W. Huang, J. Y. Kuo, Z. Yu, X. He, and B. Lei. Cross-modal transfer learning for hep-2 cell classification based on deep residual network. In *2017 IEEE International Symposium on Multimedia (ISM)*, pages 465–468, Dec 2017. 2
- [27] Y. Li and L. Shen. A deep residual inception network for hep-2 cell classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 12–20. Springer International Publishing, 2017. 2
- [28] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440. IEEE Computer Society, 2015. 3
- [29] mrgloom. Awesome Semantic Segmentation. Github, 2017. <https://github.com/mrgloom/awesome-semantic-segmentation>. 2
- [30] D. Parthasarathy. Classifying white blood cells with deep learning. online: <https://blog.athelas.com/classifying-white-blood-cells-with-convolutional-neural-networks-2ca6da239331>, March 2017. 2
- [31] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. In *CVPR*, 2017. 2
- [32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI (3)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 2
- [33] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017. 3, 4
- [34] A. Shpilman, D. Boikiy, M. Polyakova, D. Kudenko, A. Burakov, and E. Nadezhkina. Deep learning of cell classification using microscope images of intracellular microtubule networks. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1–6, Dec 2017. 2
- [35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2
- [36] A. Singh. River ice segmentation with deep learning. online: https://github.com/abhineet123/river_ice_segmentation, April 2019. 4
- [37] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3899–3908, 2016. 2
- [38] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 11 2016. 3
- [39] J. yves Bouquet. Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000. 4
- [40] L. Zhang, L. Lu, I. Nogues, R. M. Summers, S. Liu, and J. Yao. Deeppap: Deep convolutional networks for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, 21(6):1633–1643, Nov 2017. 2