Improving Image Segmentation of Docked Boats using Genetic Programming

Nick Aksamit
Department of Computer Science
Brock University
St. Catharines, Canada
na16dg@brocku.ca

I. Introduction

Image segmentation is the task of identifying and segregating different parts of an image based on some wanted features. There are many applications for such a task, but one of which may be crucial is in medical practice. Currently, a variety of deep learning techniques such as convolutional neural networks, recurrent neural networks, and long short-term memory are applied for image segmentation tasks [1]. Depending on the pairings of methods used, certain challenges can be mitigated. For example, it was found that a contextual long short-term memory model produces sharp segmentation results, which can be notoriously difficult given that many classification images used for training are hand-drawn. Having sharp segmentation can be very practical in situations such as pre-surgical medical imaging, or even in the detection of pedestrians with self-driving vehicles.

A method commonly found in many image segmentation models is the use of transfer learning [2]. In the process of transfer learning, a model is first trained on a large set of data slightly related to the current task. In the case of image segmentation, this may be a large set of general images that contain various animals and objects. Afterwards, the same pre-trained model is then trained on a set of data specific to the application, such as chest x-ray scans for detection of certain illnesses. It has been shown that when the model is pre-trained on data that is similar, but not exact to the specifics of the task, it gives better performance.

Similarly to deep learning, Genetic Programming (GP) has also been applied for image segmentation. Riccardo Poli proposed a strategy for creating image filters that can be used for image segmentation, all with a GP system [3]. In their work, such a model was applied to magnetic resonance images of the brain, along with x-ray coronagrams, and was compared against similar experiments performed but with the use of neural networks. It was found that GP vastly outperformed neural networks on this task, and the provided segmentation images show structures with clearly defined borders, unlike the neural network which is at times noisy.

This work also brings forward a method of image segmentation with the use of GP. It is formed as an extension of a previous work which attempted to segment boats. From that study, it was found that it is difficult for GP to correctly classify boats when they are located near a dock. Since boats are likely to be found near a dock when not in use, this may lead to issues. Thus, this study ventures into various methods about how a GP system can go about correctly segmenting docked boats.

Type	(xMin, yMin, xMax, yMax)
Training 1	(28, 55, 276, 233)
Training 2	(36, 242, 291, 430)
Testing	(996, 210, 1211, 442)

TABLE I
Coordinates of Testing and Training Regions

The rest of this study is divided into the following parts: Section II explains all the information needed to replicate the completed experiments, and gives information about the hand-drawn segmentation image, as well as the GP library used. After that, the results of all experiments is discussed and illustrated in Section III. Statistical analysis is performed to determine the superior variant of segmentation, and visual comparisons between performance images is done. Concluding is Section IV, where generalized remarks are given about the significant findings from this study. The potential areas of future work will also be provided.

II. Experimental Setup

This section provides all the information needed to replicate the results found. More specifically, information about the images, GP language and parameters, experiment types, fitness calculation, and statistical tests are all found within this section. It should be emphasized that the DEAP [4] system was used alongside Python 3.

A. Image Data

For the purposes of this study, one classification image was used and is shown as Figure 4 in Appendix A. Previous work expressed that the GP system struggled with segmentation when boats were positioned near a dock, and used the same image as here. This work is considered an extension, whereby attempts are made to improve upon previous results.

Figure 4 has many elements that express distinct features, and they go as follows. The two green rectangles designate the training regions used, along with one red rectangle which

Function	Arity	Description
ADD	2	x + y
SUB	2	x - y
MUL	2	x * y
DIV	2	x / y
NEG	1	-x
MAX	2	max(x, y)
MIN	2	min(x, y)
MeanF	3	MeanFilter(i, x, y)
MaxF	3	MaxFilter(i, x, y)
MinF	3	MinFilter(i, x, y)
EdgeF	2	EdgeFilter(x, y)
EdgePF	2	EdgePlusFilter(x, y)
EmF	2	Emboss(x, y)
EdgesF	2	EdgesFilter(x, y)
Red	1	RedChannel(x)
Blue	1	BlueChannel(x)
Green	1	GreenChannel(x)
avgRGB	1	(Red + Green + Blue) / 3

TABLE II
FUNCTION SET

Terminal	Value
Ephemeral Constant	$x = \text{rand}(-5, 5), x \in \mathbb{R}$
Ephemeral Constant	$x = \text{rand}(-10, 10), x \in \mathbb{Z}$
FilterSizeOne	1
FilterSizeThree	3
FilterSizeFive	5
FilterSizeSeven	7
FilterSizeNine	9
FilterSizeEleven	11

TABLE III
TERMINAL SET

Parameter	Value
Population Size	750
Generations	100 (or 50/50)
Crossover Rate	90%
Mutation Rate	10%
Number of Executions	15
Number of Elites	1
Min Tree Size (Init.)	2
Max Tree Size (Init.)	4
Min Tree Size (Mut.)	1
Max Tree Size (Mut.)	2
Tournament Size	3

TABLE IV
GP PARAMETERS

is the testing region. Bright purple features are known boat classification pixels, and yellow are the known dock pixels. It should be noted that both boat and dock features were hand-drawn, and so mistakes could be made due to human error.

As was previously mentioned, there are two training regions that are declared with green rectangles on Figure 4. In addition, there is a singular testing region which takes the form of a red rectangle. The coordinates of all these regions can be seen in Table I.

B. GP Language & Parameters

The choice of GP language, along with its parameters can drastically affect the quality of results obtained, along with how quickly they are achieved [5]. As such, it is important that sufficient thought is planned into these essential GP elements. Both the language and parameters will follow from what is used in the previous work on boat segmentation using GP, where these elements were studied closely to determine outcome quality. Table II and III both express the used function and terminal set, respectively. As can be seen from the elements red, blue, green, and avgRGB of the function set, RGB imagery is used instead of grayscale. Their function is deduced from their identification, where red obtains the singular value from the red colour channel, with similar functionality for blue and green on their respective channels. avgRGB, however, returns the average intensity between all three channels.

In addition to colours of the image, filters are utilized for their prospective assistance in segmenting images. There are a total of seven filters used, most of which serve a unique purpose, while some may be for detecting a certain feature of the image for easier segmentation. The mean, max, and min filters each return the mean, maximum, or minimum pixel value within a certain range, which is expressed in Table II as i. Subsequently, three filters are included for detection of edges, however each does so uniquely. These are denoted as EdgeF, EdgePF, and EdgesF. Lastly, an emboss filter is included, and is listed as EmF.

Each of the aforementioned filters also takes an x and y integer as an argument. This is due to the inclusion of offsets when calculating filter values. In some situations, it may not be helpful to understand the pixel value directly at the current position, but instead one that is relative. For this reason, offsets are incorporated into the function set, and use the integer values generated from an integer ephemeral constant as found in Table III.

Together with the function set, a good set of GP parameters should be used to obtain an eminent solution. Table II illustrates all of the default parameter values that remain static throughout all performed experiments. It should be noted that these values are taken from the superior values found in the previous work on image segmentation with boats. A higher crossover and lower mutation rate was preferable, along with only one elite individual per generation. Comparisons were made between different population sizes, and when image segmentation is executed with boats located near a dock, a higher population gave significantly better results. In Table II, 50/50 generations represents the number of generations when two GP systems are simultaneously executed for different segmentation tasks. This is referenced with increased detail in Section II-D.

There are some other GP parameters that remain static but are not listed in Table IV. Firstly, the parent selection technique used was tournament selection. With the tournament approach, some k individuals are chosen from the current population, and the one with the best fitness is transferred as a parent for the next generation. Following this, a one-point crossover system is used. When crossover is to be performed between two parents, a point is randomly selected on both of

Classification	Colour
True Positive	Green
True Negative	Black
False Positive	Red
False Negative	Yellow

TABLE V
PERFORMANCE IMAGE COLOURATION

the parent trees, and their subtrees at that point are swapped. Since a strongly-typed GP system is enforced, it remains true that the swapped subtrees cannot have differing root types. Lastly is mutation, where a random point is chosen on a tree, and a newly-generated subtree is generated within a certain range.

C. Fitness Function

When determining how well a program segments some image, both positive and negative classification samples can be taken from the training regions. If the genetic program solution returns a value above zero, then it is considered a positive classification, and otherwise negative. Due to the size of the training regions, it is not feasible to evaluate a solution with every pixel value due to a large computational time. Thus, only a certain predetermined number of samples are taken. Algorithm 1 expresses the described fitness function. It is worth mentioning that it is not always the case that a boat is considered a positive classification. In some of the experiments performed in this work, the dock may be considered a positive classification. Any discrepancies are covered in Section II-D.

Algorithm 1 Fitness

```
Require: x be a compiled program
  hits ← 0
  for i from xMin to xMax do
      xPos ← i
      for j from yMin to yMax do
          yPos ← j
          if x ≥ 0 and classification = True then
              hits ← hits + 1
          else if x < 0 and classification = False then
              hits ← hits + 1
          end if
      end for
  end for
  return hits
```

D. Objective and Additional Information

The objective of this work is ultimately an attempt to improve upon the previous work, which was the use of GP in segmenting images of boats. Following the conclusion of the earlier study, it was clear that GP has difficulty detecting boats when located near a dock. Thus, this work further attempts to segment boats, but with a focus on those that are located at a dock.

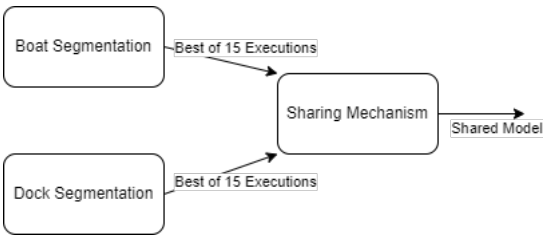


Fig. 1. Process of Experiment 3 & 4

Throughout experiments, a performance image will be shown, which is a visual indicator into how well a GP program performs. It shows the correctly and incorrectly predicted positive and negative pixels with the use of four colours. Each type of classification, and its respective colour value is seen in Table V.

A total of four experiments are completed in an attempt to correctly segment boats when a dock is nearby. The first experiment is considered a baseline, where a regular GP system is executed on the previously described parameters and language, with 100 samples of positive classifications (boat), and 100 negative (all others). Afterwards, a second experiment is performed, but a portion of samples are made from known dock pixels. This is done in order to determine if specifically sampling from dock positions assists in lowering false positive classifications overall. For the second experiment, 100 samples are taken from boat positions, 50 from the dock, and 50 from those remaining.

Two more sets of experiments are carried out following similar strategies to the first two, but are slightly different. When striving to segment boats, the dock may be classified as a boat, but it is unknown if this occurs in a reverse fashion. Thus, when segmenting for a dock, it may happen that the boats are not included as a positive. Following this intuition, segmenting for a dock may assist in segmenting for boats. Thus, in the following two experiments, two GP systems are executed, and follow a similar pattern of sampling to the first two. The third experiment will sample 100 boat data points, and 100 of those remaining. Likewise, the fourth experiment will sample 100 boat data points, 50 dock, and 50 of the rest. A flowchart of the described process can be seen as Figure 1.

Statistical testing is completed to determine if a set of results are better than another. First, the Shapiro-Wilk normality test is complete, which in turn determines the subsequent hypothesis test. If one of the found results deviate from the normal distribution, then the Mann-Whitney U test will be completed, and otherwise the Student's t-test is used. All results from statistical testing are gathered at a 95% confidence level so as to reduce any confusion from random chance.

III. RESULTS

As was discussed in Section II-D, four independent experiments are performed. In this section, the results of those experiments are addressed, and are split into two sections. One is for the prioritized sampling of the dock as compared with

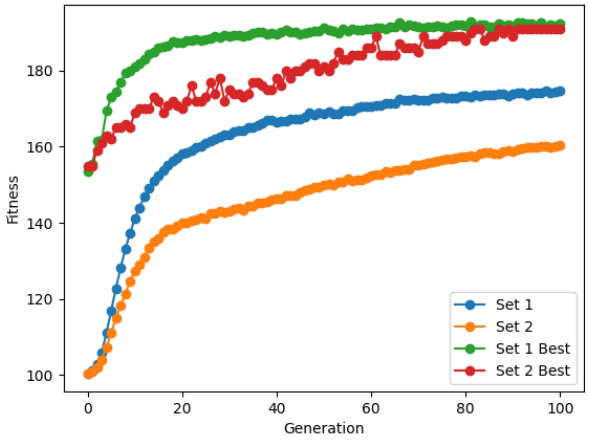


Fig. 2. Average Generational Performance (15 Executions; Experiments 1 & 2)

	Shapiro-Wilk	Student's T-Test	Mean	Std. Dev.
Set 1	0.278	0.002	43594	364.12
Set 2	0.124	0.002	**44690**	1145.85

TABLE VI
STATISTICAL RESULTS OF TESTING (EXPERIMENTS 1 & 2)

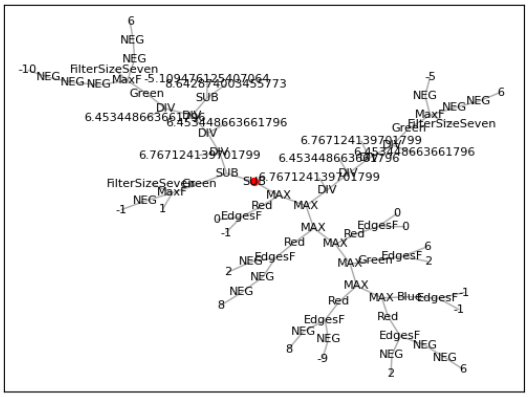


Fig. 3. Best Individual of Experiment 2

the standard sampling of boats and everything else (Section III-A), and the other is for splitting the image segmentation task into boat and dock segmentation, and then combining them afterwards (Section III-B).

A. *Prioritized Sampling*

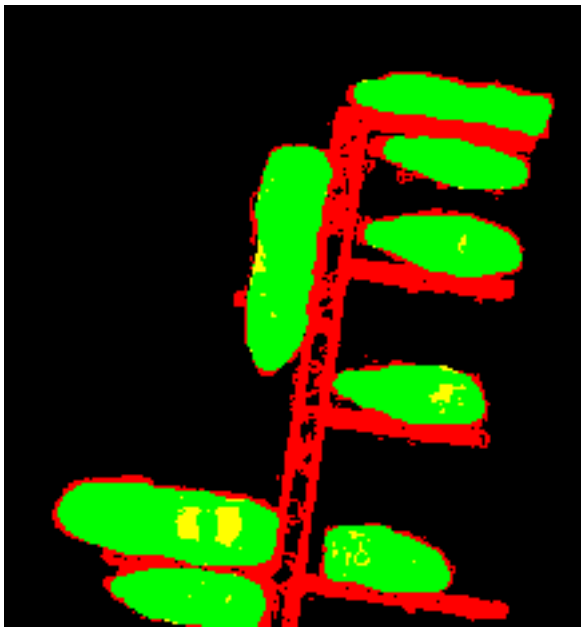
In this section, two experiments are performed to test how a prioritized sampling technique might affect the outcome of image segmentation with GP. It is known from previous work that it is difficult to classify boats when near a dock, as the dock is likely to also be classified as a false positive. Thus, instead of sampling only from known boat positions and all others, it may be of significance that known dock positions are intentionally sampled from as well. In these trials, set one performs an 100-100 sampling on boats, and everything else. Conversely, set two samples from 100 boat positions, 50 dock, and 50 remaining.

Figure 2 illustrates the average generational performance for experiment sets 1 and 2, along with their best fitness per generation, during training. It is apparent that set one overshadows that of set two, both on average and best-wise. The average outcome of set one expresses a steady increase in the earlier generations, while plateauing later on. Set two, however, maintains a slow increase, but still is not able to overtake that of set one. When glancing at the best solutions per generation, it is evident that set two is able to achieve a solution with quality near that of set one.

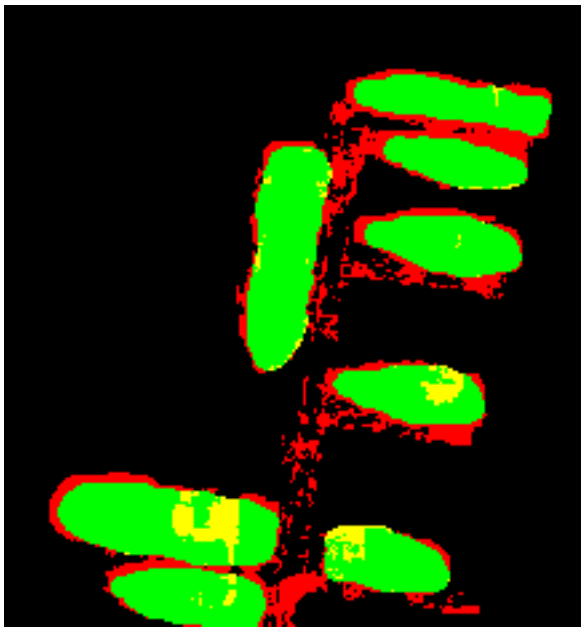
Although these training results present a superiority between the two experiments, it is important to note that because the sampling is different, so to are their fitness calculations. It is likely that set one samples from water pixels, which the GP system has already shown good performance in classification. This may grant a visual boost in performance during training, without much of an effect during testing. Consequently, set two may appear to have poor performance during training because the GP system knowingly struggles to classify dock pixels, which is sampled from 50 times each fitness calculation. Knowing these factors, it is impressive that set two is able to receive a best fitness similar to that of set one after having sampled from 50 dock pixels.

Table VI elucidates the statistical values obtained after the experiments are completed, and are from testing each best GP received over each of the 15 executions. It is clear that set two has a better mean fitness than set one over the fifteen best generated programs, but also has a greater standard deviation. The Shapiro-Wilk normality test was performed over the fifteen data points from both experiments, and it was found that both do not deviate from a normal distribution, and so the Student's t-test is performed afterwards. After concluding the t-test results, it is clear that set two is superior to set one. The notions mentioned previously are true, since the fitness calculations during training do not represent what may occur during the more generalized testing. It can be said that it is better to directly sample dock positions during the training procedure, as this may allow for a lessened chance at false positives.

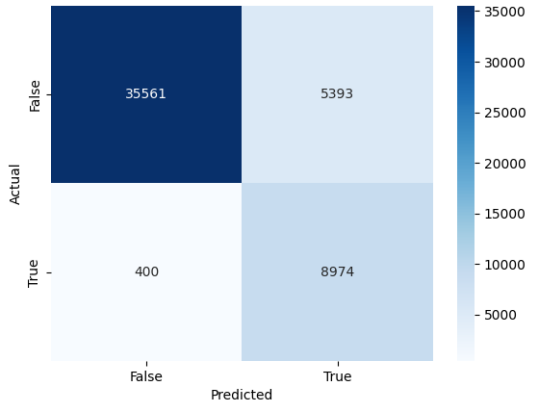
From testing, Table VII conveys the results with performance images and confusion matrices of the best individual from all fifteen executions in both experiments. As can clearly be seen, B has been refined to classify more dock locations as negative, while A struggles and so has a large amount of red pixels visible (false positives). That being said, there is a noticeable difference in some boat classifications as well. B has an increased amount of yellow pixels compared with A, meaning that the number of false negatives is larger. For the most part, B is able to correctly classify boats, and clearly



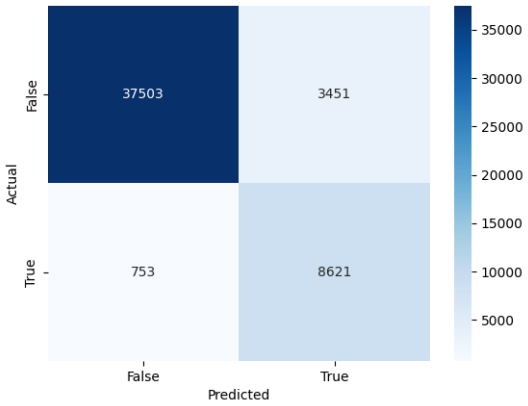
A.



B.



C.



D.

TABLE VII

TESTING RESULTS FROM EXPERIMENT SETS 1 & 2 (A AND B ARE PERFORMANCE IMAGES FROM SETS 1 AND 2, RESPECTIVELY, WHILE C AND D ARE CONFUSION MATRICES FROM SETS 1 AND 2, RESPECTIVELY)

shows an improvement over A in recognizing the dock as a true negative. The confusion matrices of both experiments (C and D) both express the results found from visually inspecting A and B. Set one has more true positives, but vastly less true negatives. This expresses that set two is much better at discriminating between a dock and boat.

A visual representation of the best program from experiment two is seen in Figure 3, where the red dot signifies the root node. As can be seen, the individual has branches with a large depth, and interestingly only uses a filter size of seven when applicable as an argument. A textual depiction of this same individual is shown directly below:

```
SUB(SUB(Green(MaxF(FilterSizeSeven, NEG
    (-1), 1)), DIV(6.767124139701799, DIV(
    DIV(DIV(Green(MaxF(FilterSizeSeven,
    NEG(NEG(NEG(-10))), NEG(NEG(6)))),
    6.453448663661796), SUB
    (-5.109476125407064,
    8.642874003455773)),
    6.453448663661796))), MAX(MAX(DIV
    (6.767124139701799, DIV(DIV(DIV(Green(
    MaxF(FilterSizeSeven, NEG(-5), NEG(NEG
    (6)))), 6.453448663661796),
    6.767124139701799), 6.453448663661796)
    ), MAX(MAX(MAX(Green(EdgesF(6, 2)),
    MAX(MAX(Blue(EdgesF(-1, -1)), Red(
    EdgesF(NEG(NEG(6)), NEG(2)))), Red(
    EdgesF(NEG(-9), NEG(8))))), Red(EdgesF
    (0, 0))), Red(EdgesF(NEG(2), NEG(NEG
    (8)))))), Red(EdgesF(0, -1))))
```

In the previous work on image segmentation on boats, it was shown that the outcome of a GP system is of higher quality when more pixel samples are taken during training. Too few, and the individuals are not able to properly distinguish positive from negative. Another disadvantage is that too many, and the computational time becomes immense and is not feasible for such a segmentation task. Thus, a trade-off arises between computational time and outcome quality, similarly to the number of individuals in a given population.

This work expresses another intuition. Not only is it important to sample as many pixels without excessively rising the computational time, it is also essential that the diversity of sampled pixels are taken into account, even with the known negatives. If only sampling from boat pixels, and all others, then there is a likelihood that dock pixels are not adequately sampled from. From Figure 4, it is clear that a majority of the negative pixels are water and not a dock. This would give GP an unfair probabilistic chance at recognizing the difference between the three main elements of the training region (boat, dock, water). However, when forcing GP to sample from each of the individual components, it performs much better than without. A drawback to this notion is the need to determine ahead of time what the individual components are of an image segmentation task. This may be an area of future work that should be taken into account.

B. Split-Shared Segmentation

Experiments three and four interpret a new approach to segmenting an image to obtain boat locations. It is known from previous work that segmenting for a boat is difficult when a dock is nearby, but it is unknown if the reverse is true. Thus, a model can be created where a GP system is executed some number of times to segment for a boat, and also for a dock. The best outcomes of both execution tasks are paired together where a sharing method is enforced. If the boat is classified as positive, and the dock is classified as negative, then a positive outcome is received since it is known that the pixel must be from a boat. Likewise, if the boat is classified as negative, and the dock as positive, then a negative outcome is accepted since the pixel must be from the dock. If both are negative then the outcome is negative, and if both are positive then the conflict management approach taken is to assume negative, since it is not possible for a pixel to be both a dock and boat.

Table VIII display the performance images of experiment sets three and four for the boat and dock segmentation tasks, and then the result of the shared approach. Observably, set three does not perform as well with differentiating between a boat and dock in comparison with set four. Nonetheless, both experiments overall performed poorly in discriminating between a dock and boat when the dock is considered a positive classification (column B). It appears that the GP models are much better at correctly identifying boats with the new sampling technique, but do not work as well segmenting for a dock. It also gives notice that segmenting for a dock is a difficult task, much like segmenting for boats that are placed nearby a dock. This debunks the previous notion that segmenting for a dock may be easier than segmenting for boats.

It is clearly seen in the C column of Table VIII that this newly defined shared mechanism performs very poorly in segmenting for boats. There is a large amount of yellow pixels where the boats are located, which denote a false negative. It seems that for a majority of the task at hand, negative classifications are made. It can also be seen that the dock is correctly classified as negative more times than their corresponding performance images in column A. Thus, using this method grants better true negative results, but does not work well with correctly classifying boats, as the amount of true positive pixels (green) in comparison with the results found in Table VII is lacking.

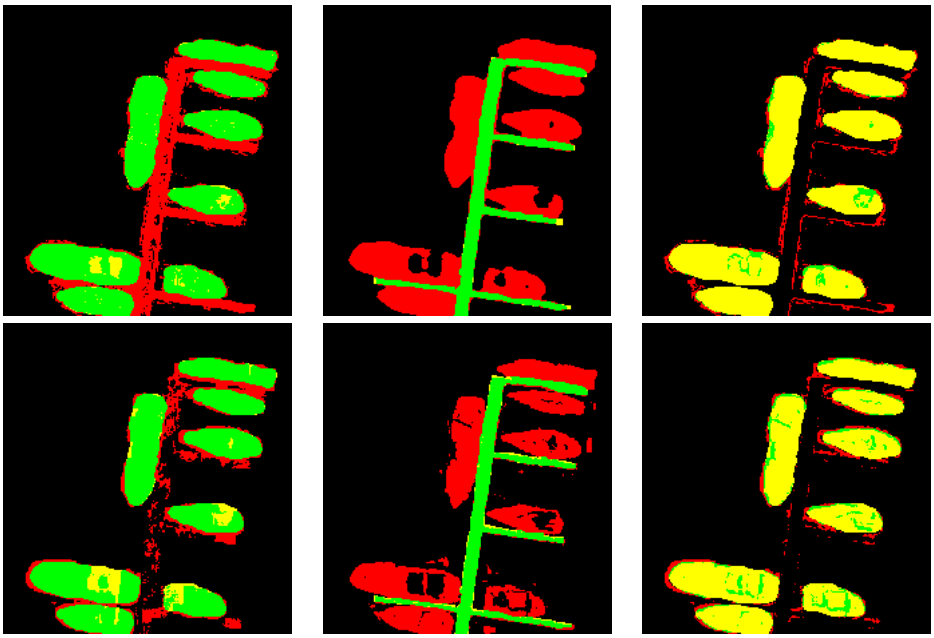
This experiment gives further insights into previous intuitions that are now known to be false. It can be seen that while segmenting for a docked boats may be difficult, so is it to segment for a dock with boats. Not only that, but it appeared GP had a more difficult time segmenting for a dock. Thus, it cannot be recommended that this procedure be performed until certain improvements are made, as the results do not express quality. Further study towards this method should be in the sharing mechanism. With better conflict management between the outcome of the two GP models, it may be possible to obtain better results. It should be noted that when using this sharing technique, the dock was more accurately classified as negative. Perhaps this could be used as an advantage in more work to come.

IV. Conclusion

From a previous work, it was shown that segmenting for boats in open water is an easily-achieved task, since a majority of the image would be blue, with a likelihood for white boats. However, this does not extend into docked boats. Given the image used in this work (Figure 4), GP struggles to correctly classify between a dock and boat. As such, this work concentrates on the improvement of GP image segmentation for docked boats.

Two different methods are attempted to achieve the task of improving segmentation. First, it was demonstrated that the quantity of samples during training has an effect on the quality of the outcome. With more samples, the GP outcome usually fares better. However, that study made no mention of the quality of samples. In this work, it was clear that the dock is an issue during segmentation, and so instead of 100 boat samples and 100 samples of other locations, 50 samples of dock pixels and 50 remaining pixels were used for correct negative classification. This gives a clear emphasis on ensuring GP is trained on a dock as well as boats, since a majority of the negative portion of training is water.

During training, it appeared that using this new sampling technique did not provide as well of a result on average. In reality however, the lower average fitness expresses a true difficulty in classification of dock pixels. It is likely that when not sampling directly from boats, fitness appears inflated because a majority of samples probabilistically come from water



A B C

TABLE VIII
TESTING RESULTS FROM EXPERIMENT SETS 3 & 4 (FROM LEFT TO RIGHT GOES BOAT, DOCK, AND SHARED PERFORMANCE IMAGE; TOP FROM SET 3 AND BOTTOM FROM SET 4)

pixels, which the GP system is already sufficient in correctly discriminating. Preeminence is shown since best fitness from the newly devised sampling reaches levels similar to that of the older, even with a worse average fitness, and difficult samples. After statistical testing, it is found that prioritized sampling of docks for negative classifications leads to a better outcome overall, and the respective confusion matrices express many less false positives. A challenge with this method is that it must be known what is challenging for a GP system to segment, and to also hand-draw the classification pixels. For future work, creating an automated system may fare better.

The next method of improving segmentation follows the notion that while it is difficult to segment for boats when they are docked, it may not be true that segmenting for a dock that has boats is just as challenging. Thus, one could use the segmentation for docks to improve segmenting for boats. After the experiments were complete, it was found that segmenting for docked boats is a less demanding task than the reverse, and therefore confirming the falseness of the previous intuition. From the illustrated performance images (Table VIII), it is clear that a large portion of the boats are falsely classified as positive, compared with the dock in column A which is not nearly as red. In turn, a sharing mechanism was used to determine how pixels should be classified after receiving results from both dock and boat segmentation. From column C in Table VIII, it is clear that this system performed a poor job at recognizing true positive classifications. For this reason, it is not recommended that this approach be used unless modifications are made.

In this work, two methods are proposed for improving the segmentation of docked boats in a GP system. It has been shown that not only do the number of samples matter, but also the quality of the obtained samples. A prioritized sampling method works well in advancing the quality of segmentation, but requires knowledge as to what samples should be prioritized. In future work, this process should be automated. Additionally, another procedure splits segmentation for one object into a two-part segmentation task, which is then combined using a sharing technique. This procedure did not prove to work well, as it improved classification of dock pixels, to the detriment of positive boat pixels. Future work in this area should consider improving the sharing mechanism.

ACRONYMS

GP Genetic Programming

REFERENCES

[1] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[2] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264, IGI global, 2010.

[3] R. Poli, "Genetic programming for feature detection and image segmentation," in *AISB Workshop on Evolutionary Computing*, pp. 110–125, Springer, 1996.

[4] F.-A. Fortin, F.-M. De Rainville, M.-A. G. Gardner, M. Parizeau, and C. Gagné, "Deap: Evolutionary algorithms made easy," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2171–2175, 2012.

[5] E. B. De Lima, G. L. Pappa, J. M. de Almeida, M. A. Gonçalves, and W. Meira, "Tuning genetic programming parameters with factorial designs," in *IEEE congress on evolutionary computation*, pp. 1–8, IEEE, 2010.

APPENDIX A
CLASSIFICATION IMAGE

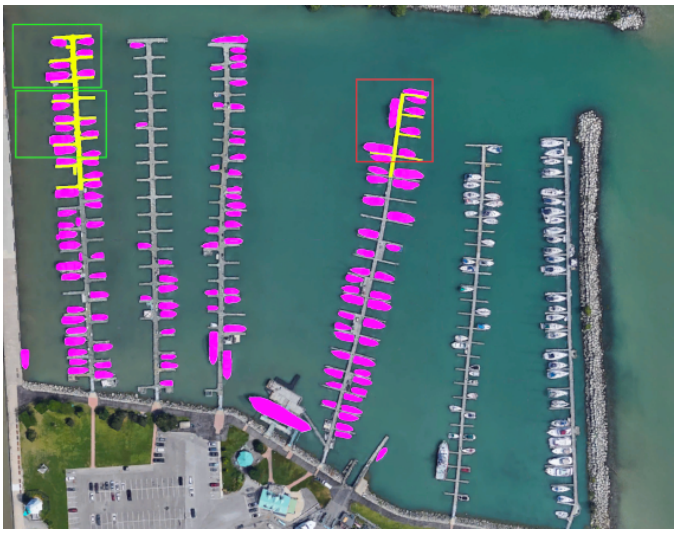


Fig. 4. Image for Segmentation of Docked Boats