An Investigation of using Simulated Data for Machine Learning

Using DeepTrack to simulate Meiofauna with the aim of segmenting experimental images

Master's thesis in Complex Adaptive Systems

ERIK BROBACK

DEPARTMENT OF PHYSICS

CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

An Investigation of using Simulated Data for Machine Learning

Using DeepTrack to simulate Meiofauna with the aim of segmenting experimental images

ERIK BROBACK

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Soft Matter Lab
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

An Investigation of using Simulated Data for Machine Learning
Using DeepTrack to simulate Meiofauna with the aim of segmenting experimental images
Erik Broback

Supervisor: Daniela Pérez Guerrero, Department of Physics, University Gothenburg
Examiner: Giovanni Volpe, Department of Physics, University of Gothenburg

Master's Thesis 2025
Department of Physics
Division of Division name
Soft Matter Lab
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 2666

Cover: Three simulated specimen, two Nodosaria and one Nematode, overlaid on an image from of a real sample. See chapter 3 for information on how the image was made.

Typeset in LaTeX
Gothenburg, Sweden 2025

An Investigation of using Simulated Data for Machine Learning
Using DeepTrack to simulate Meiofauna with the aim of segmenting experimental images
ERIK BROBACK
Department of Physics
Chalmers University of Technology

Abstract

Monitoring meiofauna is an effective way of assessing the effects of pollution in an environment. The current method of manually extracting data from samples is, however, very time consuming. If one instead uses a machine learning model for image recognition and segmentation a lot of this manual work can be automated. Furthermore, if one can use simulated data for training the model then new models can be created more easily. This work therefore investigates the possibility of simulating two groups of meiofauna, Nematodes and Nodosarias, by developing its own simulations and measuring the performance of a model trained on them. The found results shows promise both in terms of results and methodology. The findings also highlight the fact that simulations are not necessarily easily created and require new effort for every new group that should be simulated.

Keywords: Simulated Data, Segmentation, DeepTrack, Meiofauna

Acknowledgements

I would like to thank my supervisor Daniela Pérez Guerrero for her help and support during the project. I would also like to thank my examiner Giovanni Volpe for his guidance and helpful directions for the work. Lastly, I extend my thanks to the whole group at the Soft Matter Lab for their moral and technical support this spring.

Erik Broback, Gothenburg, 22nd June 2025

Contents

List of Figures

List of Tables

1

Introduction

Today, for good and bad, synthetic chemicals are a part of everyone's life. In 2019, there were around 100,000 substances containing synthetic chemicals on the EU market [1]. Included in this group are chemicals which may adversely affect human health and biodiversity, yet not all are well-understood or closely investigated [2]. As pollution is threatening one in eight species on our planet and all four European regional seas suffer from large-scale contamination problems [3], there is an apparent need for more monitoring and further investigation of the effect of pollution in our environment. To this end the CONTRAST Project, a collaboration between 10 European universities thereamong the University of Gothenburg, seeks a greater understanding of the effect of *chemicals of emerging concern* (CECs) on the marine environments of Europe and to develop new methods for monitoring as a means to achieve this [3].

A naïve attempt at measuring pollution would be to use some physico-chemical variable (such as the concentration of specific contaminants) or other abiotic variable. This, however, is not effective. A contaminant may very well be present in high enough concentration to have an effect but not large enough concentration to be detected with chemical or physical measures [4]. The presence of a contaminant alone does not necessarily translate to pollution either, as it is a measure of the biological/ecological response to the contaminant/contamination [5], [6]. This makes biological indicators the preferred method of monitoring environmental pollution, especially living organisms since they integrate all parts of the ecosystem through their various responses and adaptions.

The success of measuring biological impact is of course dependent on many factors, among them the choice of target faunal group [7]. From this perspective meiofauna has several compelling qualities that are shown in table 1.1. What constitutes meiofauna is not strictly defined, but is generally held to include lifeforms in the approximate range from 50 µm to 500 µm [8]–[10].

1.1 Meiofauna

Having established the value of monitoring meiofauna, the question instead becomes what exactly one should attempt to monitor. An optimal solution would of course be to monitor all meiofauna, yet as this project investigates new methods for detection, a delimitation of the scope is necessary for achieving progress within the time limits of the project. This project has therefore been chosen to be focused on investigate two groups of meiofauna.

Table 1.1: Attributes that makes meiofauna interesting to monitor as way to monitor pollution [8].

Attribute	Effect
Small size	Allows for small physical samples and controlled experiments in the lab
Ubiquitous distribution	Meiofauna is present in any environment that provides a source of organic carbon
High abundance	Generally a large number of individuals present which ensures statistic significance to samples
Short generation times	Most species have a short life cycle which ensures that the current population reflects current conditions

The first group is Nematodes. They are among the most numerically prevalent metazoans in aquatic habitats, from the pristine to the heavily polluted ones, sometimes representing 90% of all meiofauna at a location [11]. This dominance has allowed researchers to extract more robust data sets than would often be possible with other meiofauna or larger organisms and has made Nematodes a staple piece in monitoring meiofauna [10].

The second chosen group is the Nodosarias. They have a clear and simple shape, making them relatively easy to simulate. Their shape is also distinctive, making them a good benchmark for if the model is able to separate the interesting meiofauna from the background and debris.

1.1.1 Interesting Features

Each specimen tells a story, not just by existing but also by how they exist. Their size and shape can also inform us if the conditions at the site they have been sampled from are favourable, harsh or perhaps outright hostile to life. That is of course information that is necessary for a nuanced understanding of the degree of pollution in an area. It is therefore of great interest to not only count the number of specimens in a sample, but also to record some information, such as length and shape, about them.

1.2 Machine Learning and Simulation

Currently, the predominant way of collecting information about meiofauna in samples is to go through it by hand and marking and measuring the specimens one finds. An example of what a sample looks like is presented in figure 1.1, but note that this is only a small part of one image and normal sample taking results in many images. This process is therefore quite time consuming. There is also the method of barcoding the genetic material that is found in the sample. This method is obviously faster than going through it by hand, but only gives information about species variation, not frequency of a species or if the species has developed morphological differences as a consequence of the environment [12], [13].

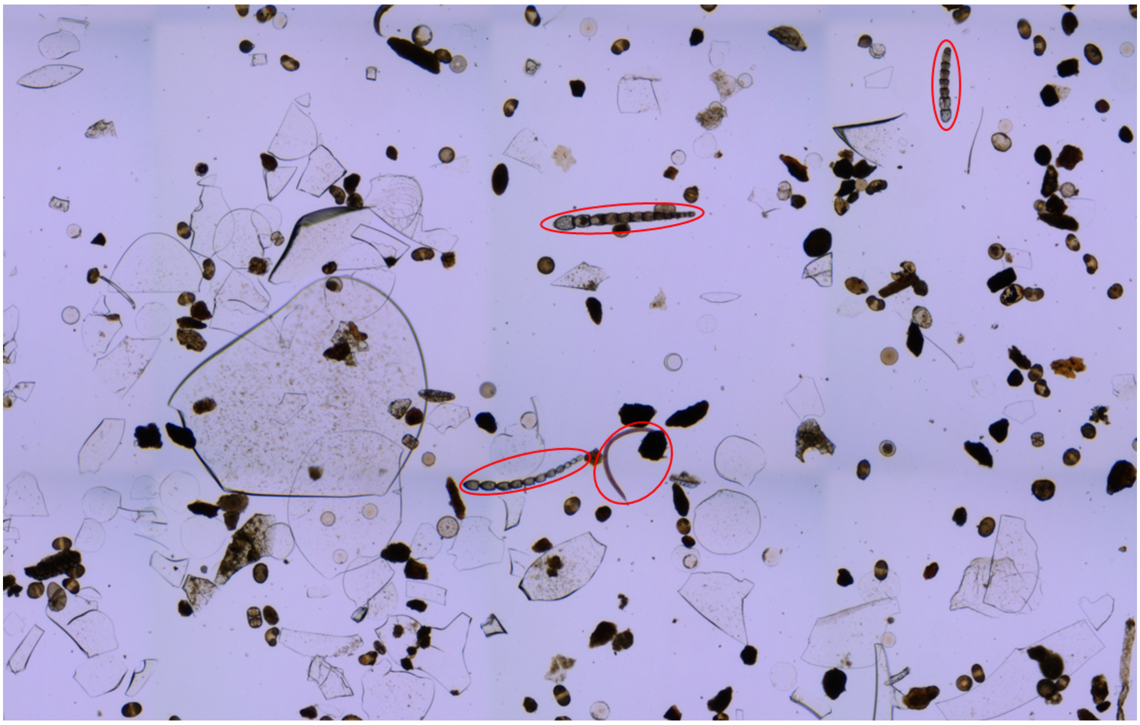


Figure 1.1: An example of what a small part of a sample looks like under a microscope. Some meiofauna specimen have been marked with red ellipses for ease of identification.

The large time needed to go trough the images manually highlights the potential benefits that could be gained by automating the process. This could be done by the use of machine learning and specifically a model trained to recognise relevant meiofauna in the sample images. This would shift the time cost from a large expense every time a new sample was taken to a, perhaps larger, single investment in training the model.

However, as mentioned in section 1.1.1, the model needs to be able to both count the frequency of a species and determine the existence of morphological differences. This makes the problem not just a question of simple detection and classification, but one of segmentation where the model returns a decision of which, if any, species a pixel belongs to for every pixel in the input image. By segmenting the shape of the specimens, the model can be used generally as it enables further measurements of various attributes to be chosen by the needs of the relevant task.

When training a machine learning model to learn to recognise objects in images, relatively large amounts of data is normally needed. As established earlier in this chapter, meiofauna is well-suited as the subject of sample taking. However, for this task where the problem is one of supervised learning the model also needs labels for its training data, in this case classification of the relevant meiofauna. Since it is a segmentation task the model must be given a classification of every pixel for the input images. Doing this preparation of the data by hand would be both labour- and time-consuming. We therefore propose that simulated data can be used instead of real data.

By simulating images, we solve the issue of manual classification as we already have access to all of the information in the images. This means that we can easily create pairs of images of meiofauna and the corresponding classification of all the pixels.

1.3 Purpose of This Investigation

The use of machine learning in order to detect and classify meiofauna has been investigated before and with good results [14], [15]. These investigations have, however, either focused on basic classification and location with real data [15] or explored simulation with simple shapes and backgrounds [14]. This work is therefore focused on two main aspects; segmentation of meiofauna as well as the use of simulated data for training of the model, in an attempt to identify and segment meiofauna in complex conditions.

2

Theory

This chapter aims to introduce concepts required for a greater understanding of the process of the project that is laid out in chapter 3. These topics are partitioned into two sections, one introducing the interesting features of the two groups of meiofauna that were chosen to be the focus of this project and one giving a brief introduction to the field of machine learning as it pertains to this project.

2.1 Meiofauna Species of Interest

As this work is intended to study the feasibility of using simulated data, the scope was limited to two groups of meiofauna. This number was judged to be enough to give insight into the possibility of using simulated data while still being possible to investigate in depth under the constraints of the project. The section aims to introduce the two groups for a greater understanding of the description of the simulations in chapter 3.

Note that all samples used in this project have been prepared by staining their contents. This means that living material will appear in a shade of pink rather than it natural colour. This is especially clear for the Nematodes and is present in some Nodosaria. Nodosaria is however mostly made up of shells, which are not living material, and are therefore not stained.

2.1.1 Nodosaria

Nodosaria is a group of organism with shells arranged on a line. Figure 2.1 shows an example of what a Nodosaria specimen can look like in the sample.

2.1.2 Nematode

Nematodes, or roundworms, are a phylum of organisms that can be found in almost all types of marine ecosystems. They therefore serve as an excellent species to monitor. Their appearance is relatively consistent across the phylum, usually exhibiting an elongated shape closely resembling to the worms many are familiar with. Figure 2.2 shows an example of what a nematode specimen can look like in the sample, stained with pink.

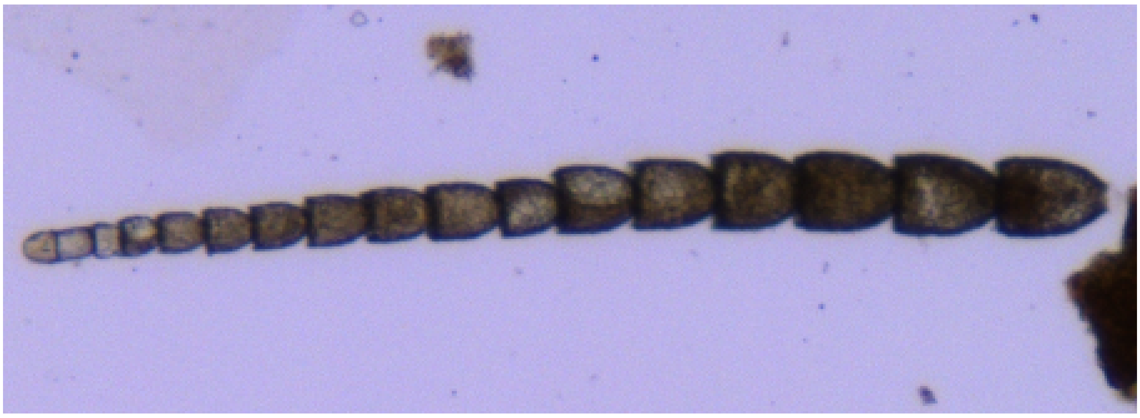


Figure 2.1: An example of how the nodosarias can look within the sample.

Figure 2.2: An example of how the nematodes look within the sample. The fading effect that can be seen is a result of merging multiple shots of the sample into one image.

2.2 Machine Learning and Image Recognition

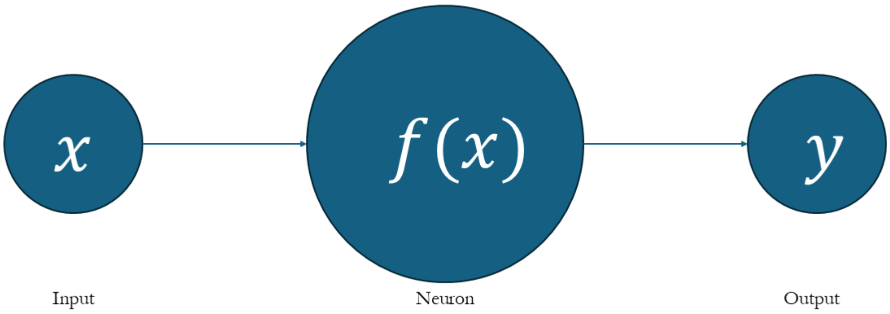
Though this work aims to be able to train a machine learning model to perform a task, the work has been mostly focused on developing the simulations and production of training data rather than developing and fine-tuning the the machine learning model and techniques. Nonetheless, a rudimentary understanding of neural networks and the specific model (U-net) used in the project does give a better understanding of the project as a whole. This section therefore serves as an introduction to the concepts of machine learning before introducing the specifics of the U-net in subsection 2.2.1.

The basic idea of machine learning is to approximate a complex relation between two (or more) variables by using a lot of small simple relations, so it is natural to begin by examining this smallest building block, the *neuron*. In its simplest form the neuron takes one input, x, and outputs the result of some linear function, $y(x) = mx + b$ [16]. From this, is it easy to expand the concept to include multiple inputs to the same neuron making the function of its output $y(\mathbf{x}) = \sum_i w_i x_i + b$. At this point we can also introduce *activation functions*, some function $g(x)$ that serves to ensure that y does not need to have linear relationship with $\mathbf{x}$, as so $y(\mathbf{x}) = g\left(\sum_i w_i x_i + b\right)$. If one further imagines a series of layers of multiple neurons where each neuron takes the outputs from the previous layer as inputs and gives its outputs as inputs to the next layer, one has what is called a *fully connected network*. These concepts are illustrated in figure 2.3.

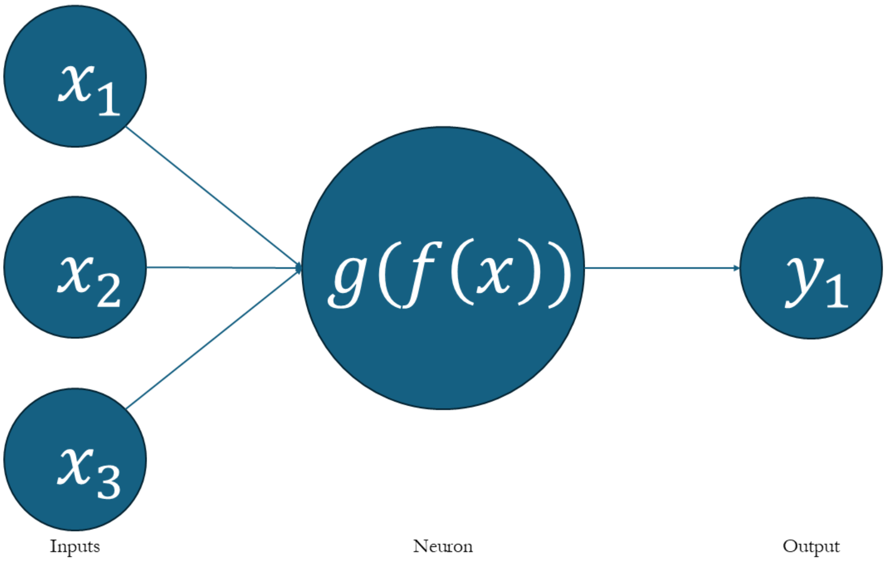
In the area of machine learning that is relevant for this project, *supervised learning*, the network is trained by giving it input–output-pairs, referred to as training data. The training is the compromised of comparing the model's output with the correct output and adjusting the model's weights to minimize the difference. Since the optimal weights are most likely slightly different between each pair, the model's weights are only adjusted a small step towards the optimal state for each pair. The model is then trained on all of the training data multiple times, and each full sweep through the training data is called an *epoch*.

As a model trains it becomes better at approximating the underlying distribution that the training data is drawn from. However, most training data is affected by some noise related to how it has been gathered. If the model is advanced enough and is trained for too long it may be able to pick up on this noise and learn to approximate it as well. This is called *overfitting*, and means that when the model is then used on another batch of data from the same distribution but where the noise has varied slightly, as noise tend to do, then data gets misinterpreted by the model that has learned to anticipate a very specific form of noise [16].

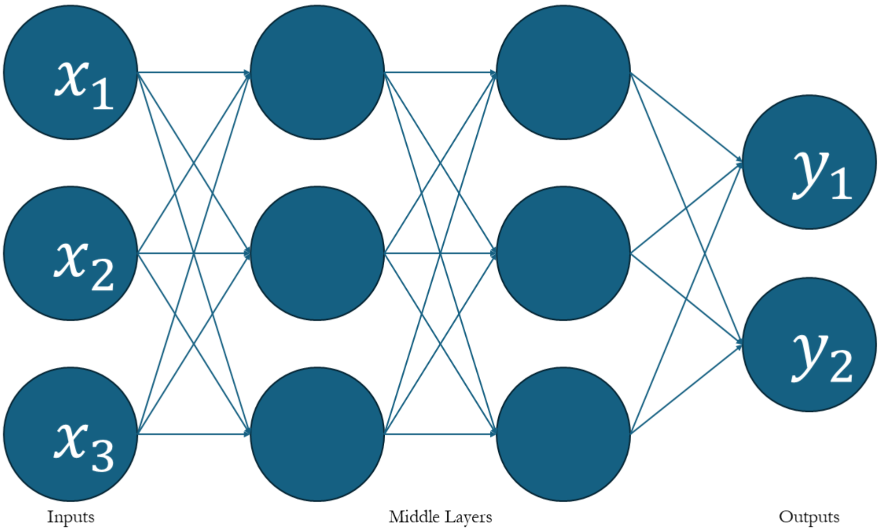
When using machine learning models to solve problems surrounding images, it is easy to view the values of the pixels as the input into the model. It is, however, important to keep in mind that unlike the case of the fully connected network, for an image spatial relations between the inputs are very important to the information that the image carries. That is, where a pixel is in the image can be of equal, or greater, importance than its value for what the image portrays. This is true for the pixel's global position, where it is in the image as a whole, but even more important for the pixel's local position, where it is in regards to the pixels around it. This



(a) The basic computational building block for neural networks, the neuron.



(b) A slightly more advanced neuron, using multiple inputs and an activation function.



(c) A fully connected network. Note that in this illustration the points representing outputs also acts as neurons, in summing the incoming values according to their weights and applying an activation function.

Figure 2.3: Illustrations of increasingly advanced uses of the basic neuron.

local position is what determines whether a pixel, for example, is part of circle or of a cross. Therefore when working with images you want neurons that also take spatial relations into account, and that can recognise the same features no matter where they are in the image.

The answer to this problem is *convolutional neural networks* (CNN). In a convolutional neural network the basic building piece is the *kernel*. It looks at a piece of the image, taking the values of the pixels in that piece and much like the neuron outputs a weighted sum. It then moves a couple of pixels over and does the same thing, weighting the pixels in the same way. It repeats this process until it has covered the entire image. By combining multiple kernels, both in parallel and in succession, with downsampling layers, such as a max-pooling layers that shrink the image by only taking the highest value in an area, the model can extract both low- and high-level features and use them to solve the intended task.

Convolutional neural networks are usually ended by a fully connected neural network in order to perform simple classifications or similar tasks, e.g. is there a traffic light in this image. However, when tasked with segmentation such as in our case, the output is not just a simple set of numbers but a new image, another architecture is instead needed. This is why a U-net is used.

2.2.1 U-Net

The U-net is a neural network architecture designed for segmentation, and developed from the “fully convolutional network” [17] which is a model architecture for segmentation that only uses convolutional networks. It does this by first downsampling the input to extract high level features, and then upsampling the result and use the extracted features to output a segmentation for all pixels in the input image at once [18].

In its contracting part the U-net has in each downsampling step two 3x3 convolutional layers that, in total, doubles the number of feature channels before a 2x2 max-pooling layer. In the expanding part, each step consists of two 3x3 convolutional layers that, in total, halves the number of feature channels and a 2x2 transposed convolutional layer, that in essence do the reverse of a max-pooling layer. An important addition to the architecture of a U-net compared to the “fully convolutional network” is that the output from each step in the contracting part is concatenated to the input to the corresponding step in the expanding part [17]. This creates a structure that when illustrated looks similar in shape to the letter U, thereby its name, see figure 2.4 for more information. The addition of connections between the segments is intended to help the model retain low-level structural information in the image, such as the rough positions of objects in the image.

2.3 DeepTrack

This project has used DeepTrack for creating its simulation. DeepTrack is a Python library built to assist machine learning by creating pipelines for image processing [19]. In this project we are mostly interested in the DeepTrack’s ability to simulate objects and placing them in images in a random yet structured manner.

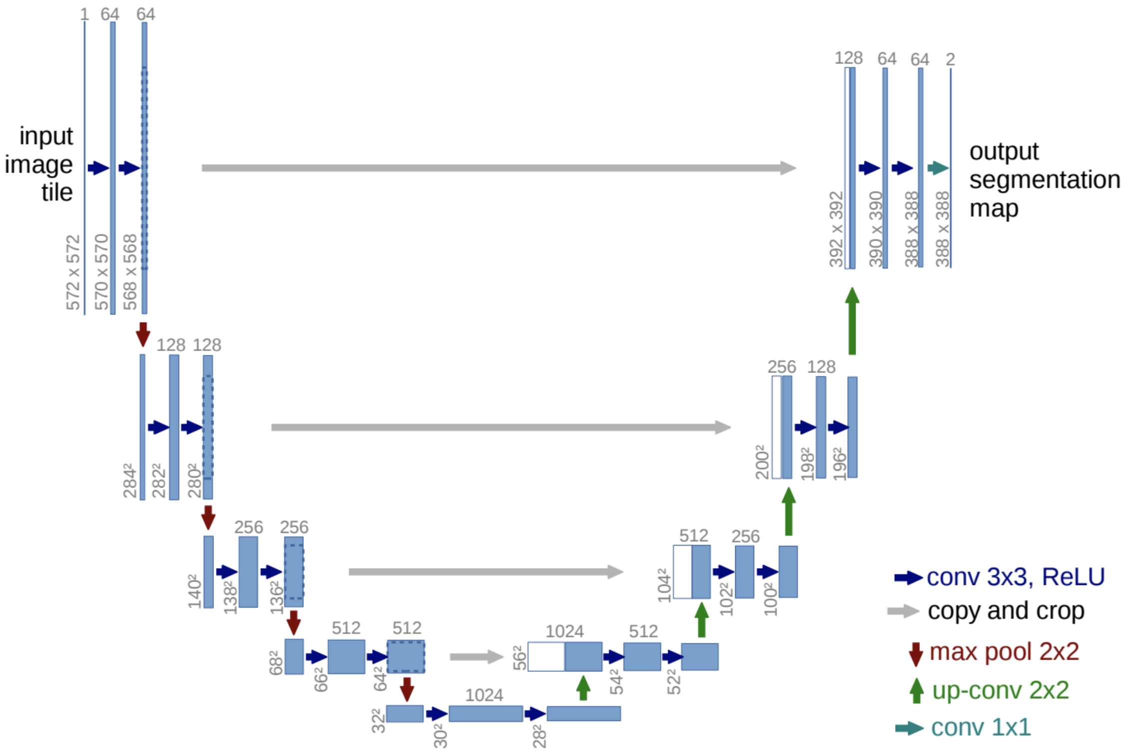


Figure 2.4: Illustration of a U-net. Note the connections between the steps as well as the contraction and expansion. Image taken from [17]. The image illustrates that the output is smaller than the input. This is a consequence of the way the kernels were designed in [17], and this effect is not present in this work.

3

Process

The project has consisted of a series of iterative changes and enhancements with the aim to improve the ability of the model to identify real meiofauna when trained on the simulated data. This chapter aims to walk you through that process, to show what worked, what did not and the challenges that still remains. Every simulated image is the result of multiple steps and parts that all affect the end result. During the project multiple pieces of the process were improved in parallel but here each topic is presented separately for ease of reading and understanding. The order of the different sections does, however, match roughly when the concept was first implemented. The GitHub repository containing the code of the project can be found here.

3.1 Basic Process

The basic process of the simulation is as follows: Set up algorithms for creating a 2D-array representing the meiofauna specimen with given parameters. Use DeepTrack's pipeline architecture to call on these algorithms for a random number (0–3) times for each meiofauna classification and place them on a larger 2D-array representing the sample. Use DeepTrack's fluorescence microscopy mode to simulate the image of the sample, inverting the light intensity to create a light background and dark objects. Originally DeepTrack's brightfield microscopy mode was used as it is the way samples are taken in reality, but since the simulations are only 2D brightfield is not suited for simulating the variance in shade that the simulations should replicate. The simulations are only generated in greyscale and the images are converted to greyscale before being entered into the model. This decision was made on the grounds that it should be easier to create a functioning model/test if a model works on the problem with only one input channel.

Using the simulated image as well as the known information about the simulated specimens, a binary mask for each category representing the successful segmentation of the meiofauna is created. The simulations and mask creation process is repeated a large number of times (300–1500). Lastly, the simulations and masks are used to train a U-net with a total of 14 layers, starting with 16 channels. This model is then evaluated by testing it on real images of the relevant meiofauna and making a qualitative judgement of how well the model is able to segment them.

3.2 The Shape of the Meiofauna

The two types of meiofauna that the project has focused on, Nodosarias and Nematodes, have both a conceptually simple shape. Though this does not necessarily translate into a simple simulation, especially when taking into account the natural randomness and diversity that can be observed in real life. It does, however, provide a good starting point for the simulation.

3.2.1 Nodosaria

Nodosarias can quite easily be seen as a series of shrinking cells in the shape of half-ellipses, and that is also where the simulation started. Measurements of real Nodosarias showed that cells were on average 93% of the size of the previous cell, with the starting cell on average being 88 by 57 µm. This was therefore chosen as the measurements of the simulation. When each cell is also allowed to be rotated slightly compared to the previous cell the result is quite similar to the silhouettes of real nodosarias. A maximum bend of $\frac{\pi}{18}$ radians (10°) in both directions was decided as a compromise between allowing larger bends of the Nodosaria as a whole and making sure that the joint between each cell looks natural. An example of this simulation can be seen in figure 3.1.

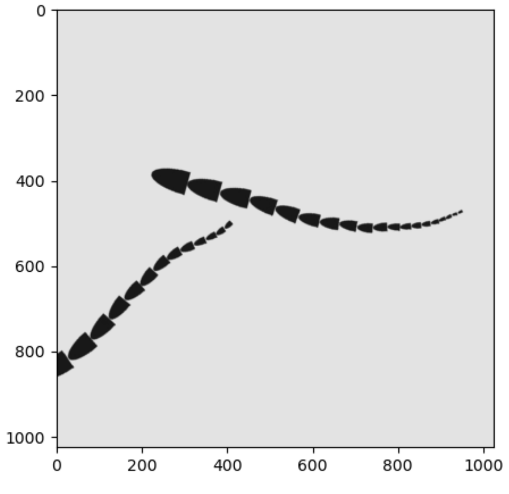


Figure 3.1: An example of an early Nodosaria simulation with two specimen. Note the independent bending between each cell as well as the consistent size decrease.

However, if one looks at real Nodosarias, one quickly realises that very few specimens are quite as orderly as the early simulations. Compare for example the two specimens in figure 3.2. One of them has very long ellipsoid cells barely connected to each other, while the other has almost circular cells pressed tightly together. A good simulation must allow for both of these situations.

The first step towards this improvement was to make the reduction rate between two cells semi-random, even allowing for later cells to be larger, as well as making the

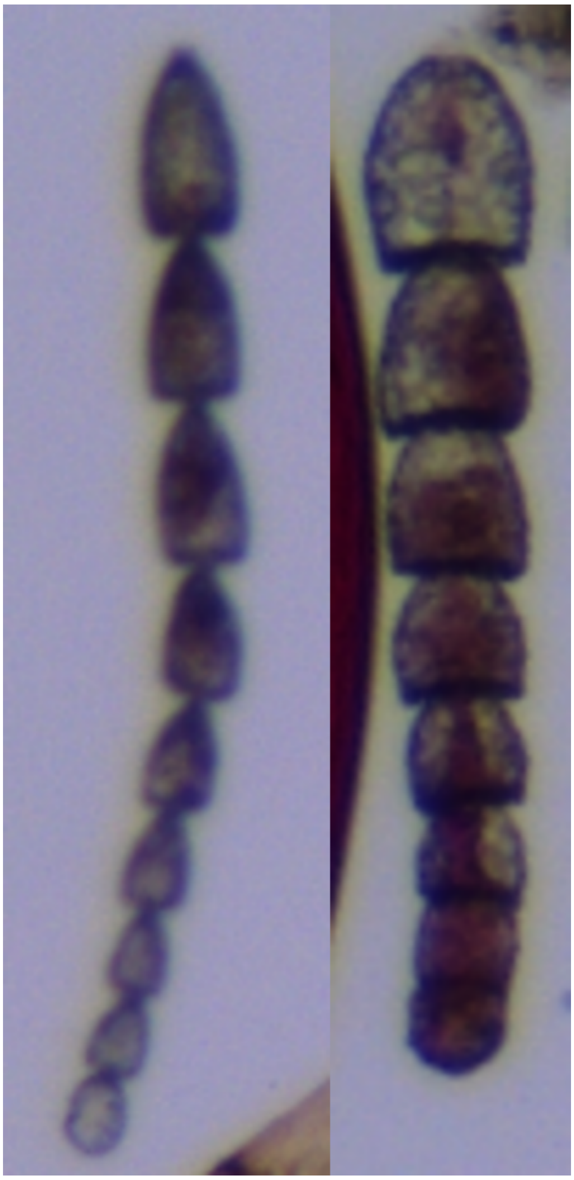


Figure 3.2: A comparison between two Nodosarias specimens. Note the differences in cell shape (ratio between height and width) and how the cells connect to each other

reduction rate for the height and width independent from each other. This allowed for a greater variance of the shapes that could be represented. However, testing showed that the independent reduction rates between height and width too often resulted in that one of the attributes shrunk to a minimum long before the other creating very unnatural shapes. As a response the reduction rates were once again tied to each other, to require them to be roughly equal but not necessarily the same. In the same vein, the size of the initial cell was also shifted to be more random by starting from a 97 by 67 µm cell and reducing it with the random reduction rate before creating the first cell.

The second step was to allow the cells to overlap. In the early simulation, the following cell connects at the very base of the previous cell, often with nothing more than a single pixel in overlap. This configurations is unlikely to appear in real specimens so the simulation was updated so that the cells were instead connected with a random amount, but no more than 20%, of the cell overlapping.

The third step was to vary the amount of allowed cell shapes further. By including the possibility for the cells to include up to 15% more of the ellipse in the cell shape, the process avoids having all the cells be perfect half-ellipses. Allowing the cell to include more than just half of the ellipses creates a shape where the widest part of the cell is not the base but instead slightly above it, a behaviour seen quite regularly in real Nodosarias. All of these improvements, combined with other changes described in the other sections of this chapter, leads to the simulation presented in figure 3.3. The image is noticeably more pixelated than the early simulation, which is because of the shift in resolution as described in section 3.5.

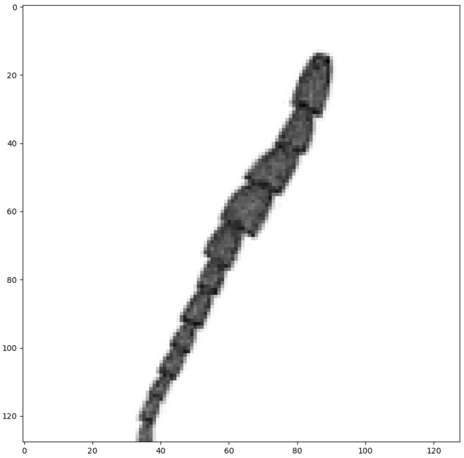


Figure 3.3: An example of the latest Nodosaria simulation.

3.2.2 Nematode

Nematodes are a type of worms, and can therefore easily be seen as just one long ellipse, contorted. This is also were the simulation began, by simulating an ellipse and giving it a bend as illustrated in figure 3.4. This implementation did, however, not allow for the bending required to accurately simulate real Nematodes. First, the simulation only allows for one bend, where the real Nematode continuously bend in different directions. Second, the implementation limits larger degrees bending and requires the beginning and the end of the Nematode to be on a line from which the rest of the Nematode bends from. This is obviously not realistic either, with real nematodes not seldom bending over themselves.

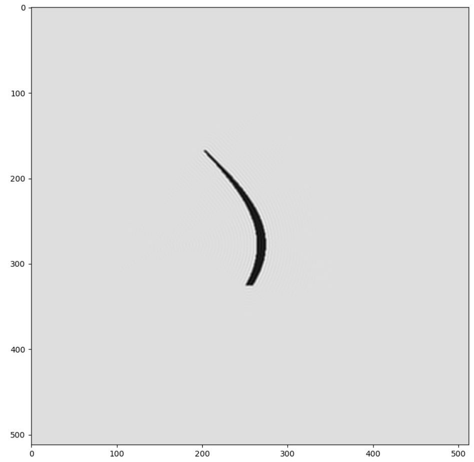


Figure 3.4: An example of an early Nematode simulation. Note the simple shape.

The solution, inspired by the implementation for the Nodosaria and the fact that the Nematodes are also segmented, was to instead build the Nematode iteratively with rectangles, each being allowed to be slightly rotated compared to the previous one, and using circles as joints and to make sure that gaps does not appear in the silhouette where the rectangles have rotated away from each other. This implementation allows for more realistic shape and bend with the right parameters (e.g. number of rectangles and allowed rotation). The parameters were shifted over the course of the project, in the end 200 rectangles were used and the allowed rotation was $\frac{\pi}{16}$ radians (≈11°). Lastly, the simulations allowed for Nematodes with a length between 500 µm and 2 mm and correlated width between 20 and 80 µm.
The simulation does, however, still have room for improvement. Both in the overall construction, which will be explored further in chapter 4.2.2, but also in fine-tuning the parameters of the construction. An example of such an improvement would be to tie the number of segments to the length of the Nematode.

3.3 Simulating Background and Debris

The first simulations consisted simply of a random number of simulated specimen on a blank background. It is an intuitive start but has a couple of drawbacks that were soon discovered. Because the background is completely white (which it is not in the real images) and because there is nothing but the relevant specimen in the simulated images it caused the model to simply label most things as relevant, see figure 3.5 for the result of the model on real image.

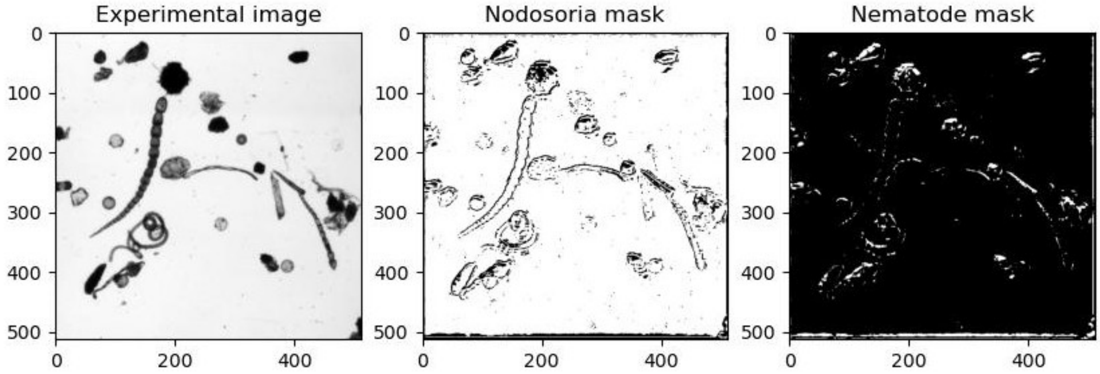


Figure 3.5: The model's output when trained on a simple simulation with white background. Note how the classifies almost everything as either Nodosaria or as Nematode.

The next step was to start introducing debris, objects in the simulation that were not to be classified as either of the relevant meiofauna by the model, in the simulation. This was done by simply creating circular objects and subjecting them to heavy elastic transformation and noise. The result was a model that was quite good at picking up objects it had not been trained to consider as debris and marking them as Nodosaria. See figure 3.6 for an example of the result of this model on a real image.

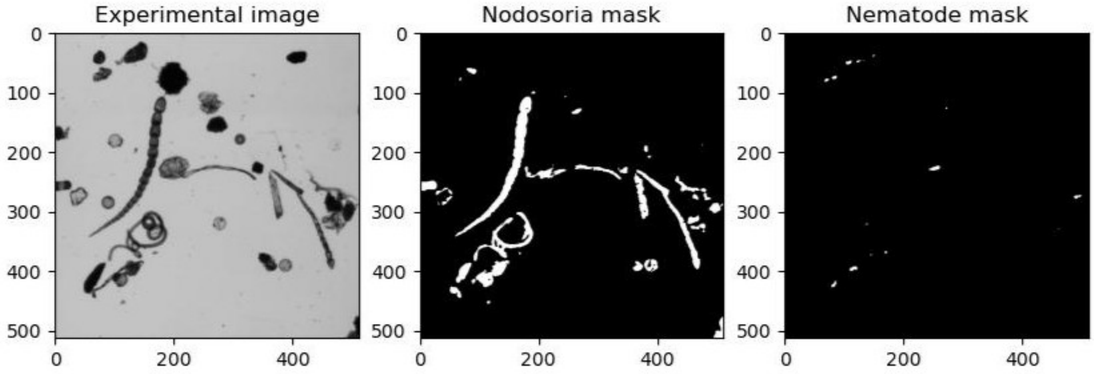


Figure 3.6: The model's output when trained on simulation that includes simple debris. Note how the model manage to identify the shapes and ignores those that are similar to the simple debris in the simulation, but does segement the debris that is not similar to the simulated debris.

This was a promising development and showed that the model could be trained to disregard debris, however it presented a new challenge; simulating all the debris and do it well enough that the model could "see" the difference between real debris and real meiofauna while still being able to generalise to unseen configurations of debris.

The found solution was to not simulate the debris, but instead to use real images. By using real images to create the background and debris a very high degree of realism could be achieved with a small level of required labour. Since the background and debris does not contain anything that should be segmented out, no time is needed to create the correct labels, as long as one make sure that the images used does not contain any of the relevant meiofauna species.

The process of inserting the simulation into a real image is relatively simple. The two images, the real and the simulated, could not simply be added together since this may cause values outside the scale or desaturation of features if the images were normalized afterwards. The images could, however, be compared on pixel by pixel basis and a new image could be created by the darkest pixel in each positions. This process is based on the idea that the samples are made up of a light background with dark objects on it. If a dark simulation is placed on a lighter piece of debris it stands to reason that the simulation should overshadow the debris and vice versa for a darker piece of debris. In perfect simulation the two objects should of course combine to create something that is darker than either of the parts, but as explored earlier less dark than the sum of the parts. Finding the equation for determining this value was not seen as prioritized in this work, and is instead left as a possible improvement for further work.

The new solution of course had its own problems. By adding very realistic background noise to the parts of the training images that the model was trained not to pick out, the demands of the simulation to accurately reproduce these effects were raised. This is clearly shown in figure 3.7, illustrating the results of a model trained shortly after the switch to real backgrounds, where the models fails to pick out any part of the relevant meiofauna. This is believed to be because the simulated specimen still had some unrealistic characteristics that made them stand out as simulated. The model then learned to identify these unrealistic features as part of its decision process, rather than basing the classification solely on their structure or other characteristics that they share with the real specimens.

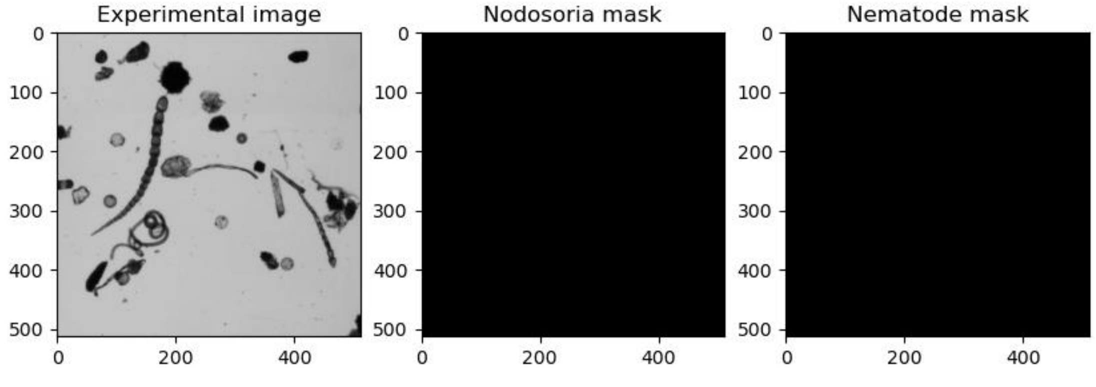


Figure 3.7: The model's output when trained on simulations on real background. Note how the model is unable to generalise to the real specimen and fails to pick out any of them.

A late addition was of the normalisation of lighting. Since different sample have different lighting, the real objects, both meiofauna and debris, shift in the shade that the model sees depending on from which batch the image comes from. The simulations on the other hand, though random, are always simulated with values within the same interval. By normalising the background images, the simulations can be tuned to have the correct shade levels for all of the background images and discerning simulated specimen becomes a little bit more difficult for the model.

Another step to making the simulation better was the development of the insertion technique. The first technique, as described above, meant that the inserted simulations always sharp, that is a clear distinction between what is and what is not the simulation. Real objects, on the other hand, cast shadows and can be slightly out of focus making their edges a little bit more blurry on a pixel level. One can relatively easily blur an image, which can be done with the simulations, but as long as the simulations are blurred before inserted they remain sharp along the edges once inserted. If the simulations are instead blurred after being inserted the entire image get blurred, misrepresenting the debris. The solution applied in this project was to first insert the simulations into the real image and blurring it, determine which pixels would be affected by the simulations by doing the same on a blank image, and inserting those pixels onto a copy of the same real image.

3.4 Creating Realistic Noise

The first simulations were solid block of colour or shade of grey, see figures 3.1 and 3.4, that were tuned to the general shade of the real specimens that they tried to represent. In order to create a bit more realism Poisson noise was applied to the simulation in order to replicate the grain that sometimes appear on photos.

Poisson noise is generated by drawing from a Poisson distribution, that is

$$P(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where k is the number of events and λ the interval [20]. This process simulates the faults that may appear when taking and working with the sample and its image [21].

This is obviously not enough to simulate the varied nature of the real meiofauna, where different specimen can vary quite heavily in colouring and pattern. “Noise” in the real world is also often more structured than the pixel-independent Poisson noise. As such, an individual approach was required for each grouping of meiofauna.

3.4.1 Noise for the Nodosaria

The first step was to create the characteristic darkening towards the edges, while still allowing for a general randomness in shade. This is achieved by the use of three layers that make up every cell. The first layer is the one that grows darker proportionally with the radius. This gives the light center and dark edges that a lot of Nodosaria exhibit, with a continuous gradient between them. Then a layer consisting of the edges of the cell is added to replicate the effect that happens when the three dimensional bodies are projected onto a two dimensional surface. Lastly a layer of random shade is placed over the entire cell, making sure that every part of the nodosaria has some colour and allowing for variation in shade. This process

is done for each cell individually, after which the entire shape is normalised to place the shade in the correct interval. After some testing, the random layer was linked to the random layers in nearby cells, thereby making the entire Nodosaria have a more uniform appearence.
The Nodosarias, like most animals, do however exhibit more variation in their colouring than simple shading gradients. In order to simulate the continuous and random variations OpenSimplex noise was used. OpenSimplex is a variation of Simplex noise, developed to be open to the public [22]. Simplex noise in turn is gradient noise with continuous change in value and without directional artifacts, and a further improvement on the more well-known Perlin noise [23]. By using OpenSimplex noise the random “patterns” within the cells could be better simulated than by using a pixel-independent noise, such as Gaussian or Poisson noise.

3.4.2 Noise for the Nematode

The Nematodes come in multiple sizes and their colouring changes depending on that, which makes simulation a bit more difficult. The larger the specimen the darker it is, but that also changes the degree to which its internal structure shines through. In essence, the centre and the edges of the Nematodes are darker than the rest. What is considered the centre is of course dependent on size and is more or less non-existent for the small specimen. This was represented in the simulation by letting the parts first be created in a darker shade before making a certain amount of pixels along the edges lighter and then once again shading back the very edges. Thereby only allowing a darker centre if the nematode was wide enough that the stretch of lighter pixels from the edges did not reach each other.
The way the nematode is constructed, with half overlapping pieces, also gives the simulation a particular noise in its appearance. This noise was left as it was and used as a substitution for any other noise. But this method does carry the disadvantage of not having immediate control over the noise, and is something that could be developed further.

3.5 Making the Shapes Realistic

The samples used in this project originally have a resolution of roughly 1.9 µm and is downsampled to roughly 3.8 µm before inputted to the model to save computational power. The simulations therefore started out with a resolution of 4 µm in order to save on memory and computational power, however as smaller and smaller details were simulated it became clear that the discrete maths that were used in the simulation didn’t work with the relatively large resolution step. The answer to this was to first do the entire simulation with a resolution of 1 µm before downsampling to the relevant resolution. This system is also scalable to different resolutions, should the parameters of the experiment change in the future.
Creating realistic shapes does not necessarily require that shapes closely follows the mathematical formulas that have been selected to approximate its real life counterparts, but rather that the shapes deviate from these abstraction in a random but still controlled manner. This was implemented by using DeepTrack’s elastic trans-

formation on the finished simulation. The transformation distorts the simulated specimen which should let the simulation more closely replicate the morphological variations found in real specimen.

3.6 Allowing for Generalisation

Despite the changes and improvements presented earlier in this chapter, the simulations were still different enough from real samples that as the model trained to identify the simulations they would "overfit". They would gradually begin to rely the unrealistic features of the simulations for their decision, and correspondingly lose the ability to recognise real Nodosarias and Nematodes as members of their respective class. The results of the penultimate version of the simulations, presented in figure 3.8, while still better than the results shown in figure 3.7, still lacked the ability to accurately pick out any real meiofauna.

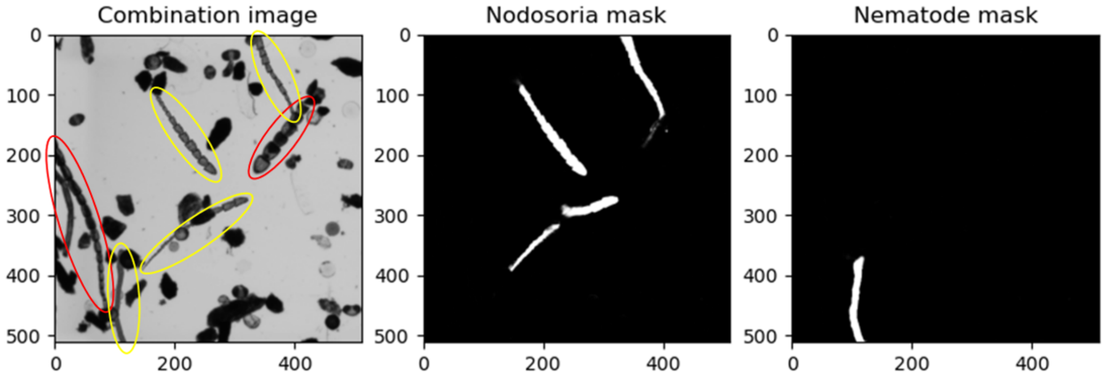


Figure 3.8: The model's output when trained on simulation incorporating all improvements, with simulated specimen in the input marked with yellow and real ones marked with red. Note how the model quite perfectly picks out the simulated nodosaria and nematodes, but fails to pick out any part of the real meiofauna in the image.

The solution to this was to prevent the model from "overfitting" or in other words to make the model less comfortable with the simulated specimen. The first step to this was a simple reduction in the amount of epochs that the model was trained for, from 150 to 100. The second step was to move the elastic transformation to after generating the mask. This means that the mask no longer was a perfect representation of the simulated specimen and the model is forced to learn to correct for the difference and not just follow the simulated specimens' appearance. The effect of this change is illustrated in figure 3.9, where a clear improvement is visible, and in figure 3.10, that shows that this improvement is not universal. So while this action does help the model it is not a silver bullet. Further action is still required and it is not clear whether that action should focus on making the simulation better or preventing the model from overfitting. Especially since there exist a limit for how much a model can be stopped from overfitting before it is unable to "fit" at all.

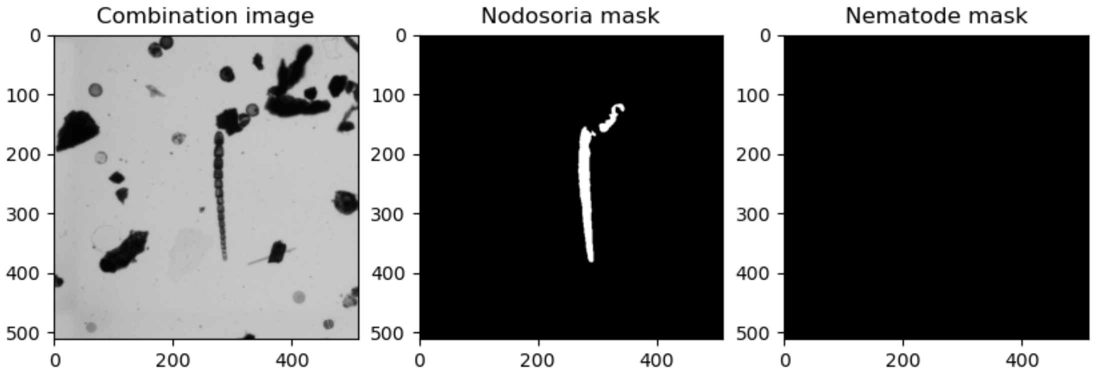


Figure 3.9: The model's output when trained on simulation incorporating all improvements and measures to prevent overfitting to simulated data. Note how the model accurately picks out the Nodosaria in the image and gives only a small false positive near its top. The result is however not universal as shown in figure 3.10.

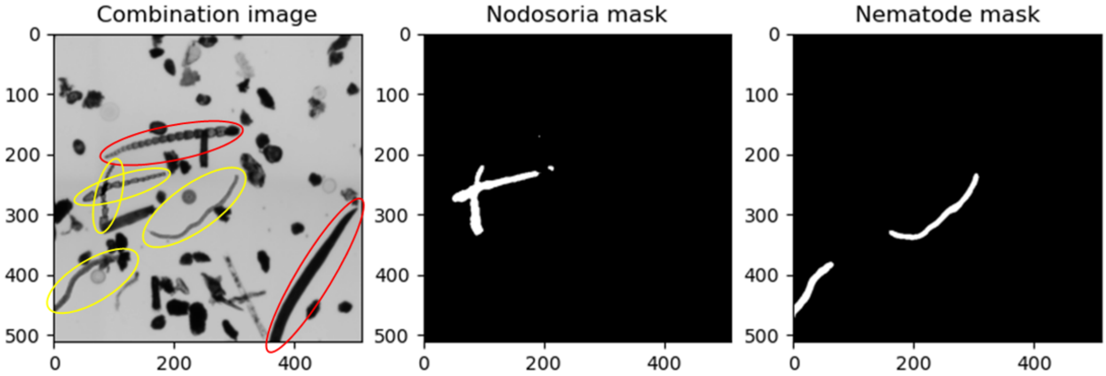


Figure 3.10: The model's output when trained on simulation incorporating all improvements and measures to prevent overfitting to simulated data. The simulated specimens in the input are marked with yellow and the real ones are marked with red. Note how the model accurately picks out all of the simulated specimens but fails to identify the real specimen. Note however that the model performs much better in 3.9.

4

Discussion and Conclusion

The project has shown clear improvements in simulation and in performance of the model. It also leaves several points from which further work can begin. This chapter aims to analyse what went well in the project, what went less well and possible directions for further work. The chapter will also analyse the overall methodology and its conditions for success.

4.1 Findings

Following the results we can see that an increase in the qualitative veracity that can be observed does increase the model's ability to recognize real meiofauna. This confirms the central premise of and shows clear promise for the concept of using simulated data for training and suggest that the issue is, perhaps not surprising, in making realistic simulation.

The results stemming from the actions to prevent the model from overfitting does, however, point to that it is not about creating simulations that are realistic enough for the model to realise that they portray the same thing as images of real meiofauna, but rather to create simulations that are so realistic that the model is unable to realise the difference. A quite big step up in the required realism that will make any attempt to use simulated data harder.

4.2 Issues and Possible Improvements

Comparing figures 3.6 and 3.7 we can observe that shifting from simulated backgrounds to real backgrounds does lower the amount of false positives but it does also heavily increase the amount of false negatives. This shift to using real images in the training is probably also the reason for the high degree realism that the model requires. Simply put, if all the data the model is trained on is simulated then the model can not use signs of simulation as a guidance in making its decisions. On the other hand, as noted in section 3.3, a solely simulated approach would require more time spent on developing the simulations of debris. However, if one reviews the results in figure 3.6, which was achieved with a much less sophisticated simulation than the results in figures 3.9 and 3.10, the needed level of veracity in a complete simulation may be lower than when simulation are placed in a real environment. It may therefore be the case that the development of passable debris simulations are more efficient than the development of very realistic meiofauna. The sheer number of different debris that is, and in the future may be, present in the samples may mean that such is not the case, but it may be worth investigation in future work.

4.2.1 Issues with the Nodosaria

The simulation of the Nodosaria has reached a level were the simulated specimens are interchangeable with real specimen for casual viewers, which rates it quite highly in terms of realism. It does however still lack the variety in appearance to replicate the whole spectrum of real Nodosaria. This has been a ongoing problem in the project as reality presents a quite wide spectrum of possibilities for any given attribute, and the simulation must be able to simulate all the values. The attributes are very rarely independent either, meaning that one attribute affects the possible range of other values in a multitude of ways. It may therefore be beneficial to split the simulation of the Nodosaria in multiple different processes each simulating a different type of the possible Nodosaria, instead of as now attempting to create a process which can capture the whole spectrum.

4.2.2 Issues with the Nematodes

From the results one can see that the model performs worse on the Nematodes than on the Nodosarias. This is most likely because the simulation of the Nematodes has been overtuned. While the process in subsection 3.2.2 does build on the structure that can be observed in the original images, see appendix B this structure is somewhat removed when the image is downsampled. The simulation is also suffering from its construction. Since the patterns of the Nematode, the darker centre and edges, are built into the building pieces the of the simulation, when the pieces are added together their overlap creates a distinct type of “noise”. This noise does not have an empirical basis in the real samples, but have been left in the simulation for two reasons: Firstly, the simulation needs some form of pixel-dependent noise to portray natural variations. The second, more important, reason is that the noise is built into the simulation by the same process that ensures that the simulation can replicate the shape of real Nematodes. Removing it while keeping any other form of internal structure of the simulation would require a new way of constructing the simulation, something that this project was unable to find. Disregarding internal structure was, however, briefly attempted as the simulation was regressed to produce a solid shade of colour throughout the entire simulation. No conclusive evidence was found, making it a possible avenue for further research but perhaps not an indication that it should be seen as a priority.

In the same way that the Nodosaria has a large variation of appearence in the real world, so too does the Nematode especially when it comes to colour variation dependent on size. This may be because larger Nematodes takes up more of the staining, or it may be for other reasons. Nonetheless, it would probably be of interest to split up the simulation into multiple parts each one responsible for creating a certain type of Nematode.

4.3 Further Work

During the project the model has been working with images of size 512 × 512 pixels, or 1024 × 1024 for the real images before they are downsampled. The real samples are however taken on a size of 10 000 × 30 000 pixels. Using such a large image size would put a considerable additional computational stress on the machine running the model. Furthermore, it is not guaranteed that the model itself can translate everything that it has learned training on small images to working with large images. At the other hand, if one instead chooses to instead crop up the samples and run each one trough the model and thereby looking at the whole sample one must make a determination how to handle the specimens that are present on multiple crops. If a specimen stretches over one or several crop lines then the model should count it once for every crop it is present on. When the total is then counted up for the sample the number will be higher than the actual number of specimen present in the samples. A functional version of the program must therefore have figured out a way to handle this. It should be noted, though, that this is only a problem if such a difference is above the accepted error. Since the measurement is made to determine the effect of pollution in an area, an exact count is probably not necessary. How precise the count needs to be is however a question for a biologist with more insight into the field than the author of this thesis.

This project has limited itself to using greyscale images, as a way to easier test the feasibility of using simulated data and U-nets for segmentation. The project shows that there is potential in using simulated data but the model struggles with picking out details and making correct decisions. A possibility to improve this would be to use colour images. Colour images carry with them more information, especially for the Nematodes as they have been stained to stand out from the background. This would however require extra work in the simulation, as the current architecture is not made for creating multi-channel output. The method of inserting simulations into real images would also need to be reviewed as the interaction between colours would be needed to be simulated more realistically than the method currently does. Should one be able to overcome these hurdles, however, the model would most likely be able to use the extra information to better distinguish between relevant meiofauna and debris.

4.4 The Approach of Simulating Data

This project was done with the aim of investigating the prospect of using simulated data for training a machine learning model to segment meiofauna in real samples. The main advantage of using simulated data to train is that one can easily create an arbitrary amount of training data, perhaps even with certain specific attributes, in a short time. Compared to the time-consuming task of creating labels for real data by hand the idea sounds appealing and would allow researchers to spend their time doing more important tasks. However, this premise only holds true if one assumes that the time it takes to design a simulation is small enough. As this project has shown, even designing two simulations is a large task and while some general insights have been gleaned most of the improvements have been done for the specific simulations.

Should one want to create a more general model, capable of segmenting a large number of classes of meiofauna, one must realise that each class is going to need an almost artisanal touch in order to function properly. Creating such a model would therefore be a time-consuming task in its own right. Further work aimed to solve the ultimate task, segmenting meiofauna in an easy way not necessarily constructing realistic simulations, should therefore spend some time investigating the options at hand.

If a functioning model based on simulation were to be created, then the human involvment would still not be over. Simulation is replication of nature, and as nature evolves so to must the simulations. However if no human eyes looks at the samples then the changes in nature may pass without notice. Already, the existence of crypto-species and little-understood early stages of meiofauna makes judgement difficult. This problem would only be exacerbated by turning over identification to machine learning models, as they would not make any notice of new changes to the species or the arrival of a hitherto unknown species and instead simply not recognise these sightings as not being relevant meiofauna and move on.

Bibliography

[1] European Enviroment Agency, “Marine messages II,” Tech. Rep. 17, 2019. DOI: 10.2800/71245.

[2] V. Dulio, B. van Bavel, E. Brorström-Lundén *et al.*, “Emerging pollutants in the eu: 10 years of norman in support of environmental policies and regulations,” *Environmental Sciences Europe*, vol. 30, no. 5, 1 2018. DOI: 10.1186/s12302-018-0135-3.

[3] The Contrast Consortium, *The contrast project.* [Online]. Available: https://www.contrastproject.eu/ (visited on 19/05/2025).

[4] G. W. Suter II, “Applicability of indicator monotoring to ecological risk assessment,” *Ecological indicators*, vol. 1, no. 2, pp. 101–112, 2001, ISSN: 1470160X. DOI: 10.1016/s1470-160x(01)00011-5.

[5] J. L. Wilhm and T. C. Dorris, “Biological parameters for water quality criteria,” *BioScience*, vol. 18, no. 6, pp. 477–481, 1968, ISSN: 00063568, 15253244. [Online]. Available: http://www.jstor.org/stable/1294272 (visited on 15/04/2025).

[6] GESAMP, “Biological indicators and their use in the measurement of the condition of the marine environment,” UNEP, Tech. Rep. 55, 1995.

[7] C. G. Moore and B. J. Bett, “The use of meiofauna in marine pollution impact assessment,” *Zoological Journal of the Linnean Society*, vol. 96, no. 3, pp. 263–280, Jul. 1989, ISSN: 0024-4082. DOI: 10.1111/j.1096-3642.1989.tb02260.x.

[8] M. Schratzberger, J. Gee, H. Rees, S. Boyd and C. Wall, “The structure and taxonomic composition of sublittoral meiofauna assemblages as an indicator of the status of marine environments,” *Journal of the Marine Biological Association of the United Kingdom*, vol. 80, no. 6, pp. 969–980, 2000. DOI: 10.1017/S0025315400003039.

[9] T. Moens, K. Sroczynska and H. Adão, “Meiofauna in a changing world,” *Ecological Indicators*, vol. 138, p. 108 769, 2022, ISSN: 1470-160X. DOI: https://doi.org/10.1016/j.ecolind.2022.108769. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470160X22002400.

[10] M. Schratzberger and J. Ingels, “Meiofauna matters: The roles of meiofauna in benthic ecosystems,” *Journal of Experimental Marine Biology and Ecology*, vol. 502, pp. 12–25, 2018, IςIMCo, the 16th International Meiofauna Conference, ISSN: 0022-0981. DOI: https://doi.org/10.1016/j.jembe.2017.01.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002209811730031X.

[11] R. Sandulli and M. De Nicola-Giudici, "Pollution effects on the structure of meiofaunal communities in the bay of naples," *Marine Pollution Bulletin*, vol. 21, no. 3, pp. 144–153, 1990, ISSN: 0025-326X. DOI: `https://doi.org/10.1016/0025-326X(90)90550-R`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/0025326X9090550R`.

[12] R. Gielings, M. Fais, D. Fontaneto *et al.*, "Dna metabarcoding methods for the study of marine benthic meiofauna: A review," *Frontiers in Marine Science*, vol. 8, 2021, ISSN: 2296-7745. DOI: `10.3389/fmars.2021.730063`. [Online]. Available: `https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2021.730063`.

[13] R. Danovaro, L. Carugati, M. Berzano *et al.*, "Implementing and innovating marine monitoring approaches for assessing marine environmental status," *Frontiers in Marine Science*, vol. 3, 2016, ISSN: 2296-7745. DOI: `10.3389/fmars.2016.00213`. [Online]. Available: `https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2016.00213`.

[14] H. Bachimanchi, M. I. M. Pinder, C. Robert *et al.*, *Deep-learning-powered data analysis in plankton ecology*, 2023. arXiv: `2309.08500 [physics.bio-ph]`. [Online]. Available: `https://arxiv.org/abs/2309.08500`.

[15] M. Plavetić, "Identification of environmentally relevant benthic foraminifera from the skagerak fjords by deep learning image modeling," M.S. thesis, Gothenburg University, Gothenburg, Jun. 2023.

[16] B. Mehlig, *Machine Learning with Neural Networks: An introduction for Scientists and Engineers.* Cambridge University Press, 2022.

[17] O. Ronneberger, P. Fischer and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: `1505.04597 [cs.CV]`.

[18] J. Long, E. Shelhamer and T. Darrell, *Fully convolutional networks for semantic segmentation*, 2015. arXiv: `1411.4038 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/1411.4038`.

[19] B. Midtvedt, S. Helgadottir, A. Argun, J. Pineda, D. Midtvedt and G. Volpe, "Quantitative digital microscopy with deep learning," *Applied Physics Reviews*, vol. 8, no. 1, p. 011 310, Feb. 2021, ISSN: 1931-9401. DOI: `10.1063/5.0034891`. eprint: `https://pubs.aip.org/aip/apr/article-pdf/doi/10.1063/5.0034891/19742219/011310\_1\_online.pdf`.

[20] E. W. Weisstein. "Poisson distribution." (), [Online]. Available: `https://mathworld.wolfram.com/PoissonDistribution.html` (visited on 05/05/2025).

[21] Wikipedia contributors. "Shot noise — Wikipedia, the free encyclopedia." (2025), [Online]. Available: `https://en.wikipedia.org/w/index.php?title=Shot_noise&oldid=1285456592` (visited on 05/05/2025).

[22] K. Spencer. "Opensimplex2." (2024), [Online]. Available: `https://github.com/KdotJPG/OpenSimplex2` (visited on 05/05/2025).

[23] K. Perlin, "Improving noise," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '02, San Antonio, Texas: Association for Computing Machinery, 2002, pp. 681–682, ISBN: 1581135211. DOI: `10.1145/566570.566636`.

A

Statement on the use of AI

As a work about the use of machine learning to solve a task, machine learning has been used through the project as laid out in chapters 2 and 3. No AI has however been used for creating this thesis or in the research beyond the uses noted above.

B

Image Gallery

Here follows some images of both real specimen and simulations for the reader that wants a greater visual understanding of the conditions.

B.1 Specimens

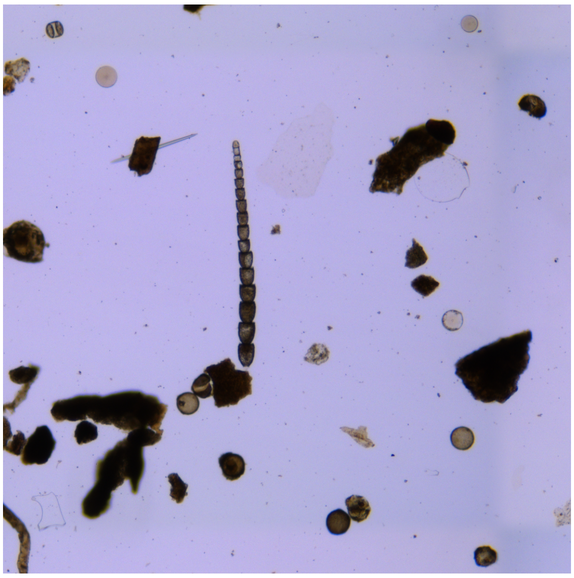


Figure B.1: A image of 1 Nodosaria in a real sample.

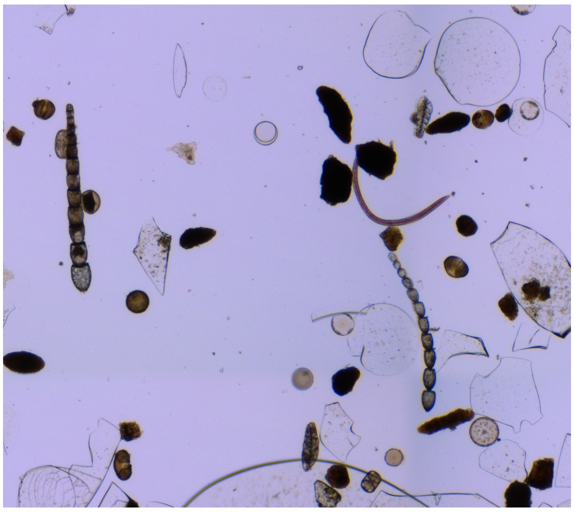


Figure B.2: A image of 2 Nodosaria and 1 Nematodes in a real sample.

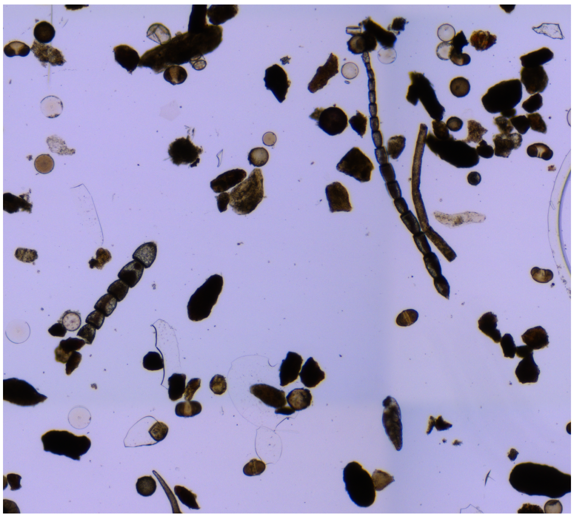


Figure B.3: A image of 2 Nodosaria in a real sample.

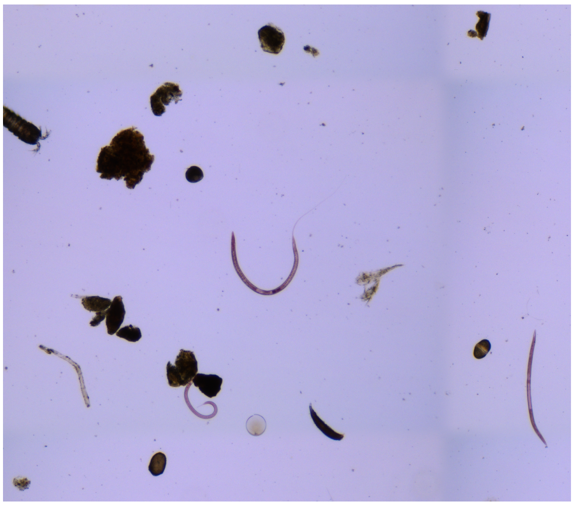


Figure B.4: A image of 3 Nematodes in a real sample.

Figure B.5: A image of 2 Nematodes in a real sample.

Figure B.6: A image of 2 Nematodes in a real sample.

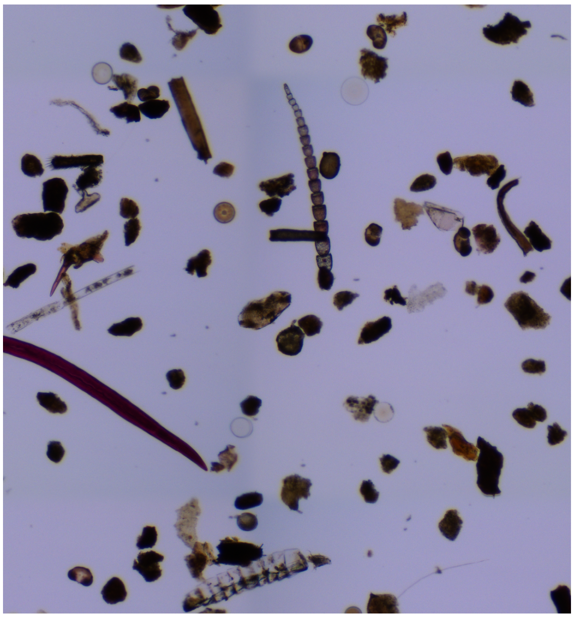


Figure B.7: A image of 1 Nodosaria and 1 Nematodes in a real sample.

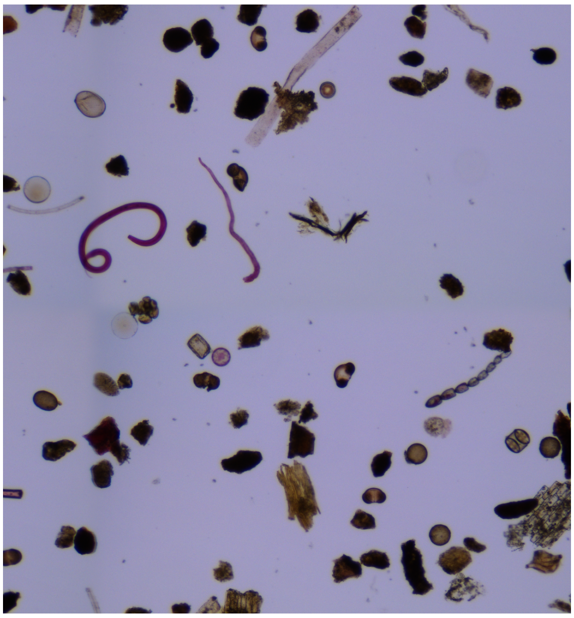


Figure B.8: A image of 1 Nodosaria and 2 Nematodes in a real sample.

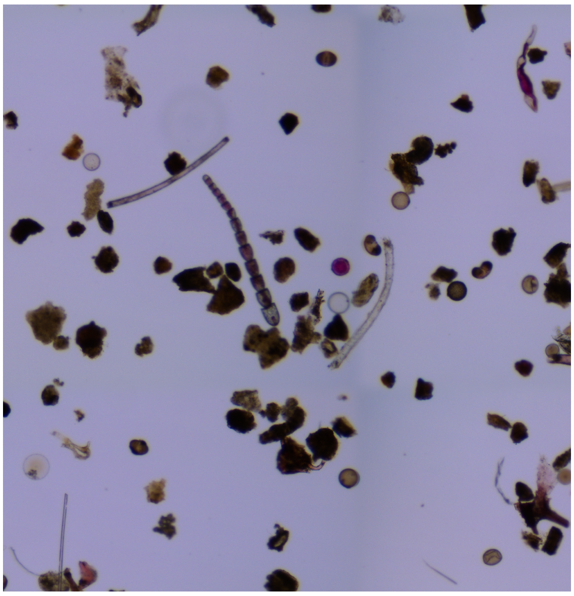


Figure B.9: A image of 1 Nodosaria in a real sample.

B.2 Simulations

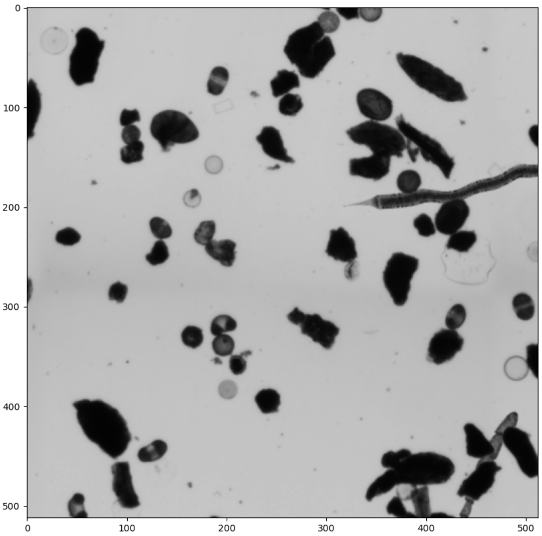


Figure B.10: A image of 1 simulated Nodosaria and 1 simulated Nemtodes on a real background.

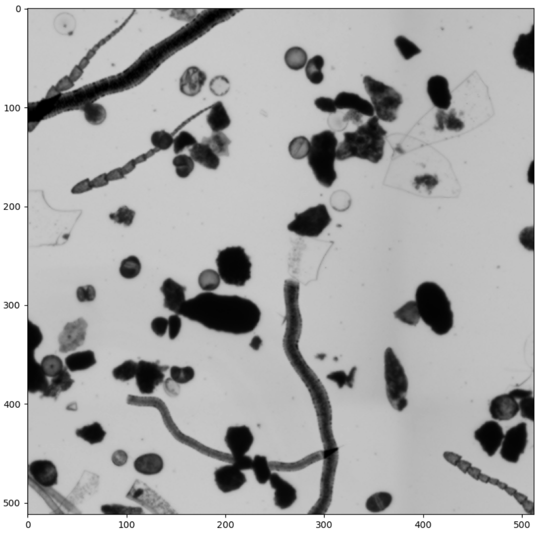


Figure B.11: A image of 3 simulated Nodosaria and 2 simulated Nemtodes on a real background.

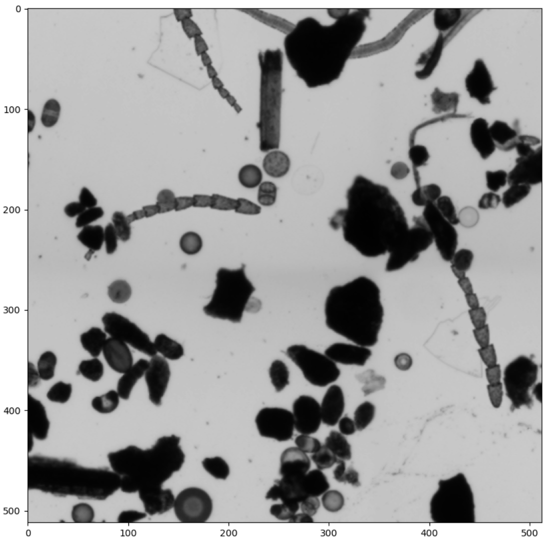


Figure B.12: A image of 3 simulated Nodosaria and 2 simulated Nemtodes on a real background.

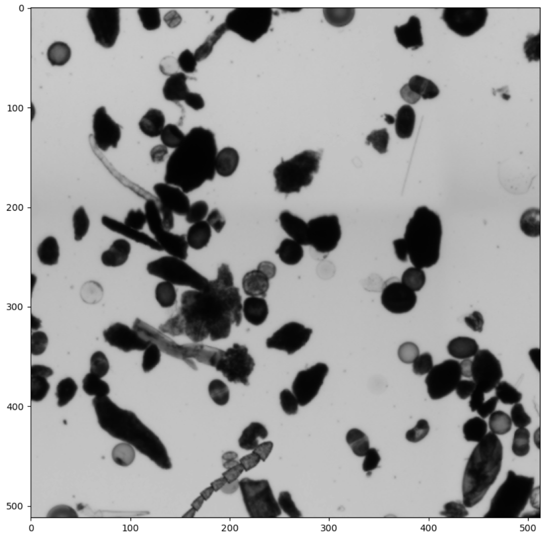


Figure B.13: A image of 1 simulated Nodosaria on a real background.

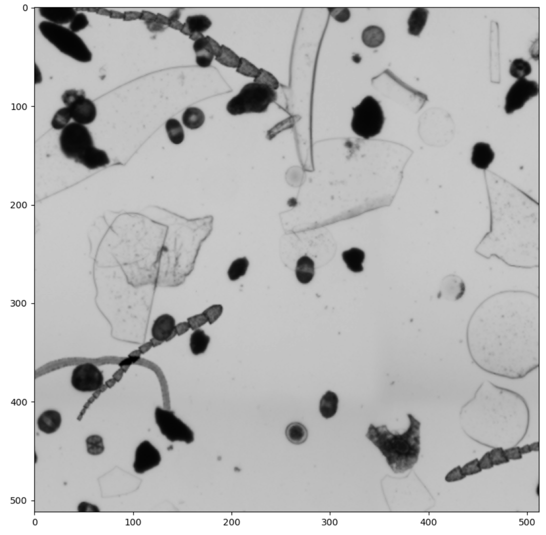


Figure B.14: A image of 3 simulated Nodosaria and 1 simulated Nemtodes on a real background.

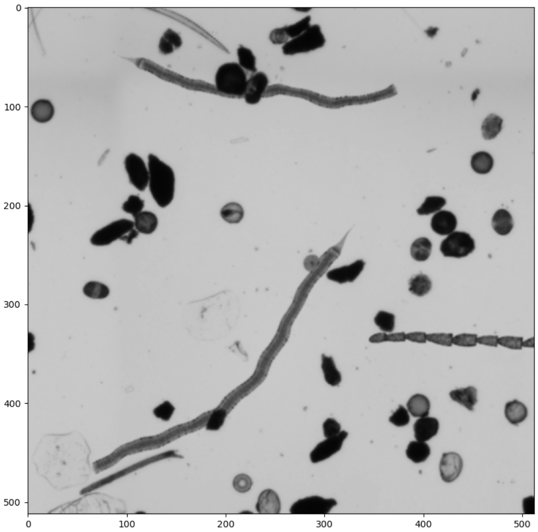


Figure B.15: A image of 1 simulated Nodosaria and 2 simulated Nemtodes on a real background.

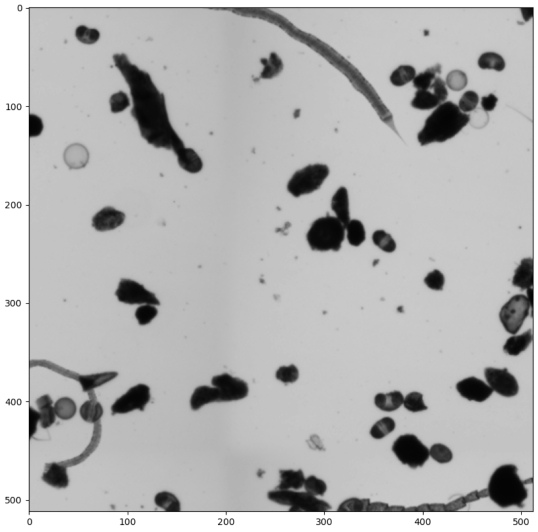


Figure B.16: A image of 1 simulated Nodosaria and 2 simulated Nemtodes on a real background.

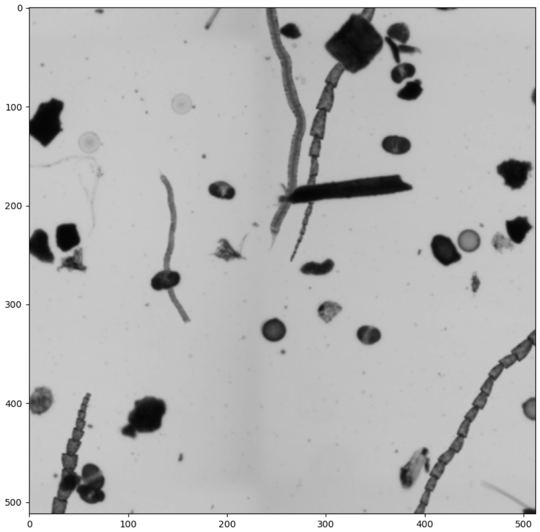


Figure B.17: A image of 3 simulated Nodosaria and 2 simulated Nemtodes on a real background.

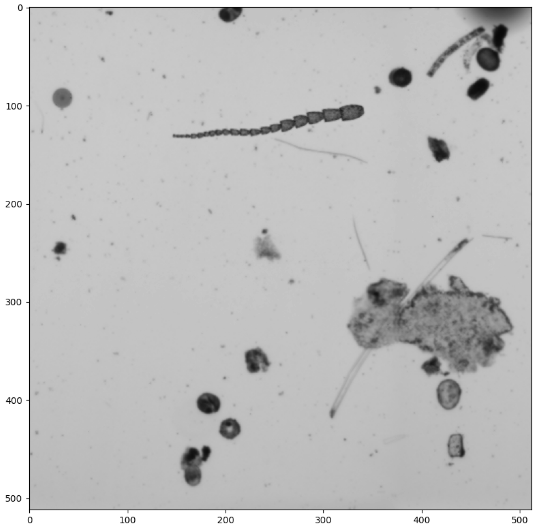


Figure B.18: A image of 1 simulated Nodosaria on a real background.

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
UNVERSITY OF GOTHENBURG
Gothenburg, Sweden
www.chalmers.se

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY