Improving the Performance of Object Counting Using Training Images in the Frequency Domain

Dani Rogmans , Yancong Lin , Silvia Pintea
[1]TU Delft

Abstract

Convolutional Neural Networks (CNNs) have made significant strides in the field of image processing over the last decade. Different approaches have been taken and improvements have been suggested. This paper looks at a newer novelty to neural networks for image counting, which is based on single-pixel center localization instead of the traditional bounding boxes. This neural network's loss function is the weighted average Hausdorff distance, which does not only take into account the number of misclassified points but also the distance between predicted points and ground truth values. The paper aims to compare the accuracy of the single-pixel center neural network on original training images of wheat heads as compared to filtered images. The filtered images have had a band pass filter applied to them, that is constructed by looking at the average frequency of wheat heads. It filters out certain lower and higher frequencies up to a threshold, and its aim is to reduce background noise and accentuate the wheat heads. Results showed that there was no significant and attributable improvement in the performance of the object counter when trained on images with filtered frequency information. A discussion of the unexpected results then carries out, with the aim of rationalizing the insignificant improvement in performance of the neural network on filtered images. As part of the discussion and conclusion, a recommendation is also made, giving insights into determining if this single-pixel center neural network is appropriate for a given dataset of images.

1 Introduction

State-of-the-art convolutional neural networks (CNNs) for object counting and localization are trained on images that are labeled using bounding boxes. In these images, a rectangular box is placed around each object of interest. However, there are certain drawbacks associated with using bounding boxes for image labeling. For example, accurate labeling of objects in image datasets using bounding boxes takes an average of 88 seconds per image [1], making image annotation a tedious and costly process. Moreover, a bounding box around an object of interest might not always be the most appropriate labeling method; crowded, overlapping and relatively small objects could be better represented by single-pixel points that indicate their center. For example, single-pixel labels of people's heads would be more appropriate than bounding boxes when trying to estimate the crowd of a concert hall. This paper aims to test this benefit brought by center-based object locators on a dataset of images of wheat heads.

1.1 Research description

Ribera et. al in [2] proposed a neural network that uses a set of single-pixel points to denote the centers of objects in an image instead of bounding boxes. In order for the neural network to train and evaluate its loss, the distance between two sets of points in the x-y plane of an image has to be computed; one set being the set of predicted points, and the other being the set actual of points. As a loss function, [2] used a modification of the average Hausdorff distance, which they called the weighted Hausdorff distance. The average Hausdorff distance $d_{AH}(X, Y)$ between two sets of points X and Y, which this modification is based on, is defined as follows:

$$d_{AH}(X,Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y} d(x,y) + \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X} d(x,y)$$

where $|X|$ is the numer of points in X, $|Y|$ is the number of points in Y, and *d(x,y)* is the Euclidean distance between two points x and y.

This is especially useful because unlike other loss measures, the weighted Hausdorff distance does not only take into account misplaced and miscounted points, but also the distance between predicted points and actual points. The neural network based on this loss proposed by [2] was shown to outperform state-of-the-art neural networks on datasets of images of people's heads, images of pupils and aerial images of wheat fields. The objects of interest in these images only spanned a few pixels in size, and were identical in terms of orientation and color. This paper aims to test this neural network on a dataset of noisier images with larger objects, in order to evaluate its generality on more troublesome images.

Some datasets of images can pose problems such as noise, blur and overlapping elements. The Kaggle Global Wheat Detection (KGWD) dataset is a collection of images of wheat

heads that are labeled using bounding boxes [3]. These images suffer from the same aforementioned problems, such as blur caused by wind and overlapping wheat heads. The dataset was picked for the purpose of the research because it allows us to evaluate the neural network on conditions that are unlike the ones it was previously evaluated on; instead of images with small and similar objects, the wheat heads in this dataset are larger and different in size, shape and orientation. Such problems can be seen in Figure 1, which shows an example image from the dataset containing 47 wheat heads to be detected.

Figure 1: An example wheat head image from the Kaggle Gloabl Wheat Detection Dataset. It contains 47 wheat heads of different sizes and inclinations. The image is also noisy, and overlapping wheat heads can be seen.

To tackle the problem of noise, we can pre-process the data by filtering out frequencies of the input images using a frequency filter. This is done by performing a discrete Fast Fourier Transform (FFT) on the input images, then filtering out certain frequencies using a frequency mask, which will reduce noise in the image and emphasize the wheat heads. It will also make the neural network less invariant to variations in the color of the wheat heads. The potential improvements to the accuracy of a neural network based on single-pixel point labeling using frequency information has yet to be evaluated, and this is most likely due to the neural network already performing sufficiently well on unfiltered images. If it can be shown that using a frequency filter improves the accuracy of a center-based object locator, this pre-processing technique could be employed in the case where it doesn't perform well enough on unfiltered images.

1.2 Contributions

The aim of this paper is to evaluate the single-pixel label neural network on the Kaggle Global Wheat Detection dataset, first using original, unfiltered images then using images filtered in the frequency domain. The difference in the accuracy of the neural network in these two scenarios will be used to determine if using frequency information improves the accuracy of the neural network. What is evaluated is not the absolute performance of the neural network under both conditions, but the relative difference in accuracy when trained on original images as compared to filtered images. We found that contrary to expectations, applying a frequency filter to training images did not improve the performance of the neural network. There are many possible reasons, the most probable being related to properties of the Kaggle Global Wheat Dataset, which makes it unsuited for this type of neural network. The key contributions of the paper are as follows:

- The results of running an object locator based on the weighted Hausdorff distance are evaluated, using multiple metrics, on original training images and filtered images in the frequency domain.
- Rationalizations of the obtained results, and hypotheses regarding why these results were obtained.
- A modification of the weighted Hausdorff distance is suggested, that it less distance-sensitive as it takes into account the size of the object

The paper is organized in the following way: section 2 discusses related works in the field of single-pixel neural networks and frequency analysis in imaging tasks, and how our study differs. Section 3 gives a high-level overview of the methodology employed in the experiment. Section 4 goes over the process of constructing the bandpass filter and how its parameters were determined. Section 5 outlines the configuration of the neural network that was used in both scenarios, as well as the results obtained. Section 6 is a brief note regarding how reproducibility of the research is maintained throughout the experiment, and section 7 offers a discussion surrounding the results. Section 8 summarizes the method and results of the research, outlines the limitations of the project and suggests topics for future research related to the research question.

2 Related Work

The domain of single-point center-based object locators has been explored by recent studies that have shown promising results. Likewise, frequency information has improved both the accuracy and training time of neural networks in multiple studies. The use of frequency information to improve a center-based object locator has yet to be studied. This section discusses different related works, possible reasons behind this topic not being studied, and differences between related works and our contributions.

2.1 Center-based object counting

Recent advances in object counting model an object as a single point instead of a bounding box. Object locators that are based on single-point centers and have no notion of bounding boxes were pioneered [2], and this domain was later extended by neural networks that first find an object's center then regress its bounding box [4]. Both of these approaches showed results that are either inline with or superior to state-of-the-art approaches such as FasterRCNN or YOLOv3, but other studies showed that the Average Hausdorff Distance was not an effective loss function for image processing tasks

such as classification [5]. The object locator used in this research is the one implemented by [2]. It was chosen because it was consistently shown to perform well on images with small, similar elements, as opposed to [4] which was tested on multi-class images with larger objects. Unlike the past works we mention, we test the accuracy of the object locator on both original and filtered images in the frequency domain, in order to see if filtering frequencies in training images can possibly improve its accuracy. Previous implementations of center-based neural networks were most likely only tested on original images because their performance was good enough on the test set. Contrary to findings discussed in [2] and [4], the neural network showed very low accuracy when trained images from the Kaggle Global Wheat Dataset; this is later rationalized in Section 7.

2.2 Image classification in the frequency domain

Image classification tasks using Fourier analysis have been tested in many works, with notable improvements in accuracy [6; 7; 8]. FFT-based convolutional neural networks were shown to improve training time [6], and the superiority of learning in the frequency domain was demonstrated "for a variety of tasks, including classification, detection and segmentation" [7]. Notable improvements in the accuracy of a deep neural network on image processing tasks were also reported [8], using a "slicing procedure that allow the network to learn both global and local features from the frequency-domain representation of the image blocks". The research carried by [8] used the cross-entropy loss function, which does not offer the benefits of a distance-sensitive loss like the Hausdorff distance. This field is widely explored, but evaluating a center-based object locator on images in the frequency domain has yet to be done. This is most likely because the emergence of such neural networks is recent and their accuracy was evaluated on original images to benchmark against state-of-the-art approaches. Our contribution is an evaluation of the center-based object locator proposed and implemented in [2], which unlike previous studies did not result in a significant improvement in performance. In section 7, we explain why this unexpected result is not representative of the impact of Fourier analysis in image processing tasks, but rather due to the unsuitable nature of the dataset for the type of neural network used in this study.

3 Methodology

The labels of the dataset are first converted from bounding boxes to single-pixel points, which is done by taking the center pixel of the bounding box. The center of the bounding box as a point label is the most logical approach to single-pixel annotation. It also complements recent approaches to regressing the bounding box of an object from a single point, which use the center as a reference [4].

The Global Wheat Dataset includes labelled images in the form [*xmin, ymin, width, height*], with the dimensions of the bounding box around each wheat head being *xmin* + *width*, *ymin* + *height*. Taking the center of the bounding box as the single-pixel label, the data has been rearranged to *xcenter, ycenter*, with

$$x_{center} = x_{min} + (\frac{w}{2})$$

and

$$y_{center} = y_{min} + (\frac{h}{2}),$$

where h and w represent the height and width of the bounding box respectively. Figure 2 shows an example image of wheat heads being represented by both bounding boxes and single-pixel centers derived from those bounding boxes.

Figure 2: An example wheat head image, with the original bounding boxes around each wheat head in blue and their respective centers in red. Looking at the figure, using the center of the bounding box as the single-pixel denotation of a wheat head is an accurate representation.

A filter mask is then applied to training images in the frequency domain in order to generate a dataset of filtered images used to train the object locator. The thresholds for the frequency mask have to be methodically chosen; [8] prescribes "trainable frequency filters that boost discriminative components in the spectrum". Filtering based on the power spectra of wheat heads and background patches aims to reduce noise in the image and accentuate wheat heads. The goal is to exploit the difference between the frequency information of wheat heads and the frequency information of the rest of the image to improve the accuracy of the object locator.

The object locator employs the U-Net architecture [9]. The visual depiction of its architecture can be found in appendix A.1.

4 Construction of the Bandpass Filter

Both original and filtered training images are evaluated in the scope of this researched, in order to measure the potential improvements a frequency filter can bring to the accuracy of the neural network. The filtered training images will first have a Discrete Fast-Fourier Transform (FFT) applied to them, in order to represent them in the frequency domain. A frequency mask will then be applied, which will filter out both low frequencies and high frequencies up to certain thresholds. This

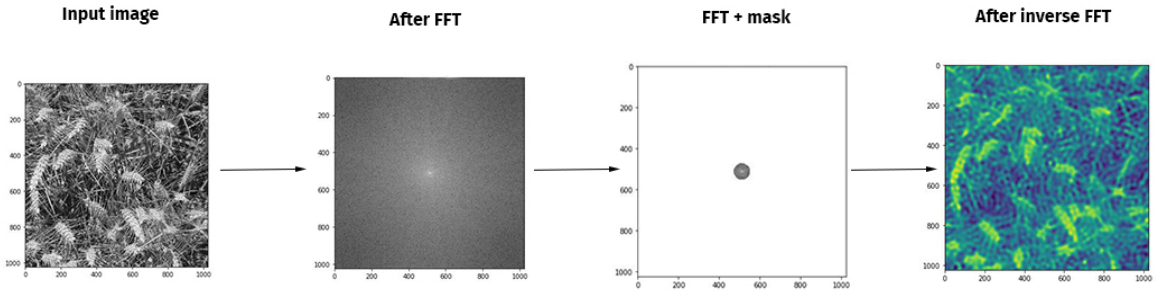

Figure 3: The process used to transform an original image into a filtered image. The input is an original image from the Kaggle Global Wheat Dataset. The image is then translated to the frequency domain using the Discrete Fast Fourier Transform. A frequency mask is applied to the resulting image, which is shown in the third step. Finally, the Inverse Fourier Transform is applied to the third image to convert it back to its original domain.

type of mask is called a band pass filter, and it is used with the goal to reduce noise in the image and accentuate wheat heads.

For the low pass filter, frequencies above 120 are cut. This value is a constant that is set from the beginning. The aim of this low pass filter is to reduce the noise in image. For the high pass filter, a meta-analysis of 115 images is used to determine an appropriate threshold. For these 115 images, the average frequency within bounding boxes, where the wheat heads reside is calculated. Also aggregated is the average frequency of the entire image. The difference of the these two is then taken, and if the difference in frequencies is above a fixed threshold of 0.2 * 10e-7, their value is included in the mask. The reason for this method instead of manually picking a parameter value for the high pass filter is because too much frequency information could be cut from the image if the threshold is too high, and having a value that is too low would not optimally reduce noise in the image. Using this newfound threshold for the high pass filter, the mask is applied to all 3373 images in the training set. These images are then converted back to their original domain using the Inverse Fourier Transform. Figure 3 shows a pipeline describing the process of transforming an input image, given that the filter parameters are already calculated.

5 Experimental Setup and Results

Although the focus of the study is related to the relative improvement in performance of the object counter with filtered images and not its absolute performance in the two scenarios, the parameters of the neural network were changed to better suit the problem. It is important to note that the configuration that is described below has been applied, with no modifications, to the object counter when run on both filtered and unfiltered data. This was done to ensure that the object locator acts as a control, and that no difference in accuracy can be attributed to different neural network configurations. Explicitly listing the parameters of the neural network also ensures that the experiment is more easily reproducible, which will be more extensively discussed in section 6.

The object locator CNN was run with the following configuration:

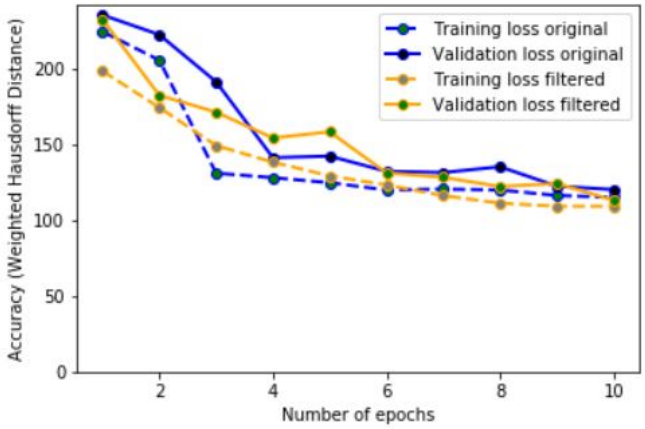

Figure 4: Training and validation loss of the neural network when trained on original and filtered images. The loss decreases over 10 epochs, but settles around 110. There is not a significant difference between the validation loss of the neural network when trained on original images as compared to filtered images.

- Dataset split: 80 percent of the images were used for training epochs, 10 percent for validation after each epoch, and 10 percent for testing.
- Image size: 512x512 (original image size being 1024x1024)
- Batch size: 8
- Epochs: 10. Early test runs showed that after 7 to 10 epochs, the drop in both training and validation loss were insignificant.
- Learning rate: 1e-3
- No data augmentation.

Figure 4 shows the average loss over each of the 10 epochs when the neural network is run on original images and filtered images. It shows that the training loss and validation loss of the neural network decreases over the course of 10 epochs when trained on unfiltered and filtered images. Both the training and validation loss decrease over the epochs in

both scenarios, but never below 110 which is a relatively high loss value. Although the losses of the neural network are lower with filtered images, the difference is minor; the last validation epoch shows a loss of 118 for original images and 113.4 for filtered images, a percentage difference of 3.98%.

Both trained models of the neural network were saved as checkpoints. After each validation epoch, the average validation loss was calculated for this epoch, and the model was only saved if the loss was the lowest so far. They were then evaluated based on the precision, recall and Root Mean Square Error (RSME) metrics. Precision is defined as

$$\frac{TP}{TP + FP}$$

where TP is the number of true positives and FP is the number of false positives. This is a measure of the ratio of positive identifications that were correct. Recall is defined as

$$\frac{TP}{TP + FN}$$

where TP is the number of true postives and FN is the number of false negatives. This is a measure of the ratio of actual positives that were correctly identified. Finally, the Root Mean Square Error (RMSE) is calculated as follows:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}|\hat{C}_i - C_i|^2}$$

where N is the number of images, C_i is the true count of wheat heads in the image, and $\hat{C}_i$ is the trained model's wheat head count estimate.

The results of evaluating the saved models on these metrics on a test subset of 10% of the dataset are shown in Table 1. The object locator was configured to consider any point within 10 pixels of the ground truth values to be considered a true positive. The table shows that the filtered training images led to very slightly better accuracy on the test set. However, the object locator performed abysmally under both conditions, and the improvement in performance is insignificant. Filtering frequencies of training images did not improve the performance of the object locator sufficiently enough, despite the parameters for the bandpass filter being carefully chosen after running an analysis over training images.

Unfortunately, the results shown in figure 4 and table 1 cannot be simply disregarded as issues related to neural network hyper-parameters. Early test runs performed with varying batch sizes, learning rates and image sizes produced similar results. The neural network's architecture was even modified by removing the 5 central layers, to no avail. A discussion listing possible reasons for both the poor absolute performance of the object locator and its lack of improvement using filtered images follows in section 7.

	Precision (%)	**Recall (%)**	**RMSE**
Original images	16.57	18.26	9.11
Filtered images	**18.11**	**18.30**	**9.02**

Table 1: The precision, recall and root mean square error of the object locator, trained on original and filtered training images, when evaluated on a test set of 337 images. The r value used is r=10, meaning that any point within 10 pixels of a ground truth point is considered a true positive. A very slight improvement in precision, recall and RMSE can be attributed to training the data on filtered images.

6 Responsible Research

When performing research on image processing tasks, reproducibility is crucial. Researchers who wish to run a convolutional neural network on filtered data in the frequency domain must be able to easily replicate the steps taken in this paper. All the steps outlined in the first three sections of this paper can be reproduced to arrive to similar results. Both the Kaggle Global Wheat dataset and the Locating Objects Without Bounding Boxes neural network are open access [2; 3].

Outlining all the configuration of the object locator in Section 3 was another step taken to ensure reproducibility. The goal was to leave as little uncertainty with regards to the parameters that were used to run the experiment.

There are still, however, some problems with the reproducibility of the experiment. For instance, running the neural network on unfiltered and filtered images over 10 epochs is very computationally expensive, and would not be able to be done were it not for the access to High Performance Computing (HPC) clusters. Without access to high-performance GPUs, the process would still be possible to replicate but would take much longer. However, because the goal of the research was to evaluate the relative performance of an object counter when run on unfiltered and filtered training images, the experiment can still be run on a simplified neural network if computational capacity is an issue.

One of the first steps taken was to convert the labels from bounding boxes to single-center points. In the spirit of promoting reproducibility, I have made the code for this step accessible in the form of a Jupyter Notebook file, and it can be found at https://github.com/dtronmans/bounding-boxes-to-centers/tree/main. The code for finding the average width and height of bounding boxes for the entire dataset, which I mention in the upcoming section, can be found at https://github.com/dtronmans/average-width-height-bbox.

Convolutional neural networks are showing steady improvement thanks to new novelties such as the single-pixel center object locator. These advancements can inevitably lead to neural networks being employed for a variety of unethical computer vision tasks, such as facial recognition in public places, tracking and surveillance. In the case of a study on human behaviour using computer vision, the scenario in which it is used should be legal, test subjects should be informed and trust should be established [10]. Most of the tasks related to human computer interaction through the use of computer vision techniques involve some degree of ethics that need to be thoroughly considered, such as unwanted discrimination of people based on certain characteristics [11] or the ethics behind creating people-centric datasets [12].

7 Discussion

Contradictory to the self-evident hypothesis that running the object locator on masked training images would improve the performance of the neural network, section 5 showed that there were no significant improvements. Two phenomena have to be discussed. The first is the poor absolute performance of the neural network when trained on original and filtered images; this will also give leads as to why the second phenomenon occurred, which is the lack of relative improvement in accuracy when the it was trained on filtered images.

7.1 Limitations brought by properties of the dataset

The most probable reason as to why the neural network performed poorly when run on both original and filtered images is the nature of the dataset. In [2], Ribera et. al ran the single-point object locator on images with small objects of similar size and shape, such as crowd heads or pupils. In the case where the neural network was tested on pictures of wheat heads, they were taken aerially by drones and therefore covered only a few pixels in height and width, making them better represented by bounding boxes. In contrast, the Global Wheat dataset contains close-up images of wheat heads of varying sizes, shapes and orientations. As an example, Figure 5 shows an example problematic wheat head, with both its original bounding box annotation and its center. It spans over 162 pixels in width and 97 pixels in height.

Figure 5: An example problematic wheat head and its original bounding box annotation. The green and purple dots, on the top-left and bottom-right corners respectively, represent possible points that hit the wheat heads but would still result in high loss when the weighted Hausdorff distance is used.

In this example, normalizing the top-right corner coordinate of the bounding box as (0, 0), the center is at (81, 49), indicated by a red dot. However, both the coordinates (5, 20) in green and (143, 76) in purple also hit the wheat head, at the top-left and bottom-right extremes respectively. If the object locator, during its training, predicts a center to be at the extreme (143, 76), this corresponds to a Euclidean distance of $\sqrt{(143-81)^2+(76-49)^2} = 68$. As the weighted Hausdorff distance employed by [2] is based on the Euclidean distance, this would lead to a high loss despite the wheat head being hit, causing the neural network to unnecessarily overadjust its weights.

To confirm that this is not only an occurrence for this particularly problematic wheat head but is reflective of the entire dataset, the average width and height of the bounding boxes were aggregated. It was found that bounding boxes had an average width of 84 pixels, and an average height of 76 pixels. Taking the center at $(84/2, 76/2) = (42, 38)$, the average maximum Euclidean distance the neural network can get while still correctly predicting a wheat head is $\sqrt{42^2+38^2} = 57$. In this scenario, the object locator would believe it has wrongly predicted the wheat heads, despite hitting it.

With filtered images, the object locator suffers from the same problem. Although filtering the images reduces background noise and emphasizes wheat heads, there is still the problem of high loss despite correctly predicting a point on the wheat head. The problem then does not reside in the neural network not being able to locate wheat heads, but in the fact that it is led to the conclusion that it did not hit wheat heads when it in fact did. It is therefore difficult to measure the improvements in performance with filtered training images, as the flawed training process of the neural network caused by overestimating errors distorts the results of the experiment.

One tentative fix for this problem would be to increase the radius r, the maximum distance in pixels to a prediction that still counts as a true positive. This won't work because of the deviation in the sizes of wheat heads. Wheat heads with smaller than average bounding boxes will count points outside the radius as true positives, and wheat heads with larger than average bounding boxes will not encompass the entire area.

There are two solutions to the problem of objects being two large for a single-pixel object locator, aside from the trivial solution to opt instead for a bounding box-based neural network like YoLo v5. The first solution is to use a loss function that is not distance-sensitive; one example would be to use cross-entropy, where we can avoid the problem of high losses despite correctly predicting a wheat head's locator. Another solution would be to modify the weighted Hausdorff distance proposed by [2], by normalizing for the size of the bounding box. The ground truth file for the dataset would then take the form *filename, counts, locations, bboxes*, where *bboxes* is the size of the bounding box around the wheat head. The average Hausdorff distance would then be modified, to be:

$$d_{AH}(X,Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y}\frac{d(x,y)}{A_x} + \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X}\frac{d(x,y)}{A_y}$$

where A_m is the area of the bounding box around a wheat head m. This would normalize the loss of the predictor according to the size of the bounding box, which is a sufficient work-around to the problem.

7.2 Limitations in the construction of the bandpass filter

Aside from the nature of the dataset, there are other reasons as to why the object locator's accuracy did not improve when run on filtered training images. These reasons are related to the construction of the band pass filter.

First, two of the parameters picked for the mask were fixed and chosen manually. The first is the threshold, which is the minimum difference between wheat head frequencies and background frequencies needed to construct the mask. This value was set at 0.2 and remained unchanged. Similarly, the exclude parameter was set at 120. This is the parameter used for the low pass filter, which removed frequencies higher than 120. As a result, the mask may have removed too much frequency information from the image, hindering the performance of the CNN. However, it is unlikely that finding a way to fine-tune these two parameters to an optimal value would lead to a much better performance. Informal runs of the neural network on filtered images using different values for the threshold and the exclude parameter were run, and these were shown to have worse results.

The exclude parameter could have been chosen by running an analysis of frequency information over a subset of training images, in a similar fashion to the method outlined in section 3. For example, the average lowest frequencies in the bounding boxes of wheat heads could be used as a guiding factor for a more methodical selection of the exclude parameter.

Finally, looking at frequency information inside bounding boxes may underestimate the difference between the frequencies of wheat heads and the frequencies of the entire image. This is because a bounding box often covers not only the wheat head, but also some background patches. Figure 5 shows such an example, where the top-right and bottom-left portions of the bounding box cover background patches. In this case, the estimated average frequency of wheat heads will be lower than its actual value, because of the background patches being taken into account. A small constant could be added to the final estimation of wheat head frequencies to counter this imbalance.

8 Conclusion and Future Work

The goal of this research was to evaluate the difference in accuracy of an object counter when trained on unfiltered training images and images with filtered frequencies. The object counter had many interesting properties, such as a novel loss function based on the Hausdorff distance, and the usage of single-pixel centers to label objects instead of bounding boxes. The novelty introduced in this neural network turned out to be the biggest downside to its performance.

8.1 Conclusion

The experiment was carried out by training the neural network on original images of wheat heads, then on filtered images of wheat heads. The filtered images were generated by running a band pass filter over them, which was constructed by looking at the difference between frequencies of wheat heads and frequencies in the rest of the image.

The results showed that there was no significant and attributable improvement in performance brought by filtered training images. The training and validation losses of the neural network were approximately the same under original and filtered conditions, and settled asymptotically at a high loss value. The two trained models were saved as checkpoints and evaluated under the precision, recall and Root Mean Square Error metrics. The filtered images performed better under all three metrics, but the improvement was very minor and the results were poor in an absolute sense under both scenarios.

It was found that the nature of the dataset was a big limiting factor to both the absolute performance of the neural network and its relative improvement using filtered images. This is because although the neural network would correctly predict a point on the wheat head during its training, its loss would be high if that point was far away from the center, despite hitting the wheat head. To confirm this, the average width and height of wheat head bounding boxes were aggregated, and it was found that they had an average width of 84 pixels and an average height of 76 pixels. In contrast, the datasets used to train the single-pixel center object locator so far consisted of images with targets that spanned only a few pixels in height and width.

The problem mentioned above hinders the performance of the neural network with both original and filtered images. Therefore, it is more accurate to say that the potential for the relative improvement of filtered images is inconclusive. The solutions proposed in section 7 aim to work around this problem; adopting them and testing the neural network under these modifications would give a clearer image of the different Fourier analysis brings.

The domain of single-pixel centers as image annotations could replace bounding boxes in a variety of scenarios, and its benefits are clear. However, neural networks relying on the center of an object should be trained on images where the target objects are small and similar, such as crowd heads in a mall. The results of the experiment therefore lead to the recommendation to check the average height and width of objects when considering training such a neural network on them. This is to ensure that the training process is not hindered by the nature of the objects in the dataset.

8.2 Future Work

Researchers trying to extend the work done on this task could follow the recommendation outlined in section 7 and find a more robust and methodical approach to picking the threshold parameter and the exclude parameter. This might not only lead to better results when running the neural network on filtered images, but the results would also be a more accurate representation of the extent to which masking frequencies in the image improve the performance of the object locator. They could also look into running the same experiment on [4], an object locator of similar nature that "uses keypoint estimations to find center points and regress to all other object properties, such as size, 3D location, orientation, and even pose".

Another improvement would be to have images with four channels in the training set. Unfiltered images contained the three Red, Green and Blue (RGB) channels, and filtered im-

ages were only compromised of one (FFT channel). Concatenating RGB with FFT would give a four channel input, which could be converted to a three channel input by adding convolutional layers. This would ensure that wheat head with higher frequencies are accentuated and that noise is reduced, without filtering out too much frequency information.

More and more approaches towards counting crowded and overlapping objects are being developed, and it is worthwhile to look at their potential potential for improvements when trained on filtered images in the frequency domain. For example, TasselNet is a recent deep convolutional neural network-based approach that "can achieve good adaptability to in-field variations via modelling the local visual characteristics of field images and regressing the local counts of maize tassels" [13]. There are also methods based on Bayesian loss for crowd count estimation with point supervision [14], whose performance could also be evaluated when run on filtered training images.

References

[1] Y. Hu, Z. Ou, X. Xu, and M. Song, "A crowdsourcing repeated annotations system for visual object detection," *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, 2019.

[2] J. Ribera, D. Guera, Y. Chen, and E. J. Delp, "Locating objects without bounding boxes," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] E. David, S. Madec, P. Sadeghi-Tehran, H. Aasen, B. Zheng, S. Liu, N. Kirchgessner, G. Ishikawa, K. Nagasawa, M. A. Badhon, and et al., "Global wheat head detection (gwhd) dataset: A large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods," *Plant Phenomics*, vol. 2020, p. 1–12, 2020.

[4] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019.

[5] O. U. Aydin, A. A. Taha, A. Hilbert, A. A. Khalil, I. Galinovic, J. B. Fiebach, D. Frey, and V. I. Madai, "On the usage of average hausdorff distance for segmentation performance assessment: Hidden bias when used for ranking," 2020.

[6] V. Nair, M. Chatterjee, N. Tavakoli, A. S. Namin, and C. Snoeyink, "Fast fourier transformation for optimizing convolutional neural networks in object recognition," *CoRR*, vol. abs/2010.04257, 2020.

[7] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," 2020.

[8] J. A. Stuchi, L. Boccato, and R. Attux, "Frequency learning for image classification," 2020.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[10] K. J. Pfisterer, J. Boger, and A. Wong, "Food for thought: Ethical considerations of user trust in computer vision," 2019.

[11] X. Ferrer, T. v. Nuenen, J. M. Such, M. Cote, and N. Criado, "Bias and discrimination in ai: A cross-disciplinary perspective," *IEEE Technology and Society Magazine*, vol. 40, p. 72–80, Jun 2021.

[12] M. Hanley, A. Khandelwal, H. Averbuch-Elor, N. Snavely, and H. Nissenbaum, "An ethical highlighter for people-centric dataset creation," 2020.

[13] H. Lu, Z. Cao, Y. Xiao, B. Zhuang, and C. Shen, "Tasselnet: Counting maize tassels in the wild via local counts regression network," 2017.

[14] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," 2019.

A Appendix

A.1 Neural Network Architecture

The neural network architecture is a U-Net, and is shown in Figure 6 below. Note that the image was taken directly from [2], and that no modifications have been done to the architecture of the neural network, except during informal test runs where the five central layers were removed.

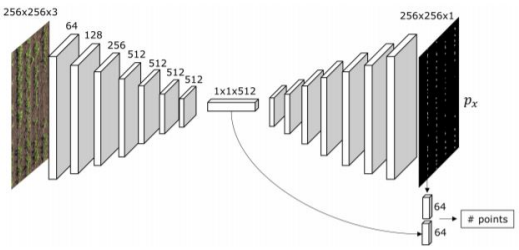

Figure 6: Architecture of the neural network used for this research, taken directly from Ribera et. al's paper "Locating Objects Without Bounding Boxes" [2]

.

A.2 Changes to code

The code for Ribera et. al's "Locating Objects Without Bounding Boxes" divides the training loss by 3, but does not do the same for the validation loss. As a result, using this neural network would give the impression that the validation loss is much higher than the training loss. This was fixed, by changing line 206 under object-locator/train.py from

iter_train.set_postfix(running_avg=f'round(running_avg.avg/3, 1)')

to

iter_train.set_postfix(running_avg=f'round(running_avg.avg, 1)')