Combining frequency information and the unsupervised W-Net model for wheat head detection

C.C.I.A. Chen, A. Lengyel, S.L. Pintea

Abstract

Wheat is a widely used ingredient for food products. To increase the production and quality of wheat, the density of 'wheat heads' in a farm can be studied. Accurately locating wheat heads in images can be challenging. A lot of work has taken place in supervised semantic segmentation, but these networks typically require large pixel-wise human-annotated labeled data. Gathering this data is tedious and labour intensive.

This paper proposes to use the novel unsupervised semantic segmentation model W-Net to solve this problem. To improve the accuracy, we investigated the influence of the frequency domain, by pre-processing the training data two different times using a custom filter, based on frequencies found in wheat heads, and a high pass filter.

The approach is evaluated on the Global Wheat Head Detection (GWHD) dataset [11]. To compare the accuracy the generated segmentations were mapped to bounding boxes based. The proposed method did not show to be able to generate competing detection compared to the baseline method associated with the GWHD dataset, but the GWHD dataset has a different measurement of truth, consisting out bounding boxes instead of segments which is in the disadvantage for the W-Net.

Pre-processing the dataset using the high pass filter did increase the intersection over union with 1,4% and the deviation of the reconstruction loss was smaller when frequency filtering was applied.

Although the object detection has a low accuracy, this study showed that some basic wheat head detection can be achieved by using the unsupervised segmentation method W-Net and the accuracy can be increased if a high pass filter is applied as pre-processing step.

1 Introduction

Wheat is a widely used ingredient for food products. To increase the production and quality of wheat, the locations of 'wheat heads' in a farm can be studied. A farmer can then asses management decisions based on the density of the wheat heads in the fields. Accurately locating the wheat heads can be challenging due to overlap of the wheat plants, blurry images and/or different colours between wheat species [27]. A variety of techniques can be applied to identify the wheat heads in images. Convolutional neural networks (CNN) has matured in recent years and revolutionized computer vision [3]. A lot of work has taken place in semantic segmentation, a supervised segmentation variant of the image segmentation problem. However, these networks typically require large pixel-wise human-annotated labelled data. Accurately pixel-wise labelling large agricultural data sets is a tedious, labour intensive and time consuming job. In this study, the novel unsupervised semantic segmentation W-Net model [2] has been used to identify wheat head locations in an image. The W-Net

model aims to remove the need for human-annotated labels by using unsupervised image segmentation. The W-Net paper claims to produce state-of-the-art results when trained on the PASCAL VOC2012 [23] dataset and tested on the Berkeley Segmentation Database (BSDS300 [16] and BSDS500 [5]) [2].

To improve the accuracy of the W-Net on detecting wheat heads, the effect of pre-processing the wheat images using the frequency domain was studied. The W-Net was trained three times: first without pre-processing the wheat data, the second time applying a custom filter and the third time applying a simple high pass filter on the training data. Several studies showed that pre-processing images in the frequency domain can make training a CNN faster and/or improve accuracy [31, 30, 19].

This study will focus on how to apply frequency information (FI) to pre-process wheat heads images. Specifically, the images will be filtered in the frequency domain based on the power spectra of the wheat heads and the background patches using the Discrete Fast Frontier Transformation (DFT) [25]. The accuracy will be measured using the W-Net model [2]. The following main question has been formulated as follows:

"How can frequency information be used to improve the accuracy of the unsupervised segmentation model W-Net, when applied to identify wheat heads in images."

2 Related work

We briefly discuss related work done using deep learning in precision agriculture, pre-processing in the frequency domain and discuss some unsupervised segmentation methods.

2.1 Deep learning in precision agriculture

Most approaches in deep learning applied on precision agriculture use supervised methods. Yamamotot et al developed a method to accurately detect intact tomato's in different growing stages. The method was based on pixel-based segmentation, blob-based segmentation and individual fruit detection. For each step, classification models were generated using the colour, shape, texture and size of the images [32]. Some methods use unsupervised learning to improve the speed the manually labelling. For example, the work of Asad and Bais [7]. It compared VGG16 and ResNet-50 for feature extraction and created segmentations using UNET and SegNet. Zhou et al proposed the "Improved ResNet" to detect broccoli heads. It has a pre-trained ResNet-50 at its core. A three layer adaptive network was added to replace the classification layers of the ResNet-50. It out performed GoogleNet, VggNet and ResNet. Zhang et al introduced an unsupervised image segmentation algorithm called Unsupervised Learning Conditional Random Field (ULCRF) to classify fruit, leaf and background. Because ULCRF is unsupervised, it cannot be told beforehand which class represents the fruit, leaf or background. In the study they found out they could use the colour feature to map the classes. The evaluation is based on the amount of pixel overlap with the ground truth [24]. By Nikbakhsh et al an unsupervised approach was introduced to segment plant leaves with complex background and combined five segmentation methods. The evaluation is based on the true or false positives (TP, FP), and true or false negatives (TN, FN) between the segmented- and ground truth area's [22]. Meyer et al tried to create a plant detection system that works using non-uniform backgrounds and applied an unsupervised clustering method

called fuzzy clustering to extract the area of interest from ExG and ExR images [17]. The papers discuss supervised and unsupervised deep learning methods to segment agricultural images. This study will use an unsupervised method and adopted the evaluation from [15] and [22].

2.2 Unsupervised segmentation

Most popular deep learning segmentation models use some kind of encoder-decoder architecture [18] and utilize features such as colour, brightness or texture over local patches [2]. Some classical clustering methods include the k-means method, Comaniciu and Meers mean-shift [9] and Gaussian mixture model [21]. Felzenszwalb and Huttenlocherâs graph-based method [13] represents images as an undirected graph with pixels representing nodes. The edges are calculated by measuring the difference between the adjacent pixels. The segments are created by taking the minimal spanning tree of the graph. The Normalized cut by Shi and Malik [10] was presented to coop with large multi-scale images in parallel, and capture both coarse and fine level details. Arbelaez et al. [6, 4] proposed a method based on contour detection by parsing any contour into a hierarchical region tree and thereby reducing the image detection problem into a contour detection problem. It presents both a contour detection and segmentation method. The original W-Net paper adopted the hierarchical grouping algorithm described by [4] as a post-processing step. The W-Net used in this paper left out this extra post-processing step.

2.3 Frequency information in deep learning

Many papers focus on compressing the memory requirements of CNN using the frequency domain [8, 29, 14]. This study focuses on the influence of pre-processing images within the frequency domain on the accuracy, rather than optimizing the memory use. Nair et al introduced a Fast Fourier Transformation-based U-Net to improve the training time and the accuracy of an object recognition model. The intersection over union score increased 55% and speed up the calculation per epoch with 30% [20]. Zhu et al evaluated several machine vision approaches for food safety. For pre-processing, they used a 2D low pass-filter, a focusing filter and a 2D Wiener filter [12]. The low-pass filter was used to remove the Additive White Gaussian noise (AWGN) created by environment and/or sensor. The focusing filter removes image blur and the 2D Wiener filter deconvolve drag effects [28]. This study will try to remove noise using a custom filter and will apply a high-frequency filter.

3 Methodology

In the first subsection, the basic workings of the W-Net model is explained. The second subsection discusses some ambiguities found in the paper and how/why this paper deviated from the W-Net architecture from the original paper. The W-Net is build for semantic segmentation, but the ground truth data uses bounding boxes. The explanation of how FFT is applied in this study is outlined in subsection 3.3. How the generated segmentations of the W-Net are mapped to bounding boxes is discussed in subsection 3.4. At last the setup of the experiments to check if pre-processing images improves the performance of the W-Net is discussed in subsection 3.6.

3.1 W-Net architecture

The main component of this study is the novel W-Net model. The W-Net model has been created by X. Xia and B. Kulis and was introduced in 2017 [2]. The model tries to solve the problem of unsupervised image segmentation. The basis of the W-Net architecture is the popular supervised image segmentation architecture U-Net [26]. The W-Net architecture uses the U-Net twice to create an auto-encoder. The first U-Net takes an input image and encodes it into a segmented image. The second U-Net tries the opposite. It takes the segmented image as input and tries to decode it into the original input image (see Figure 1). The first U-Net encodes the input image using a k-way soft segmentation. To improve the outcome of the auto-encoder, the segmentations are post-processed using two methods. First a fully connected conditional random field (CFR) is applied to smoothen the segments, secondly hierarchical merging is used to merge segments. In Figure 2 an overview of the W-Net architecture is shown. [2]

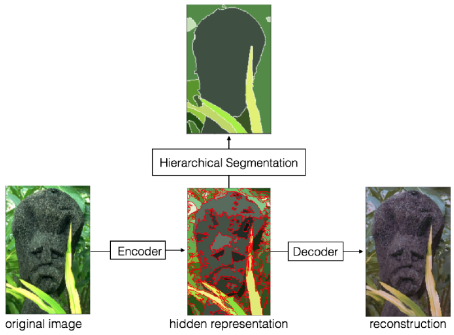


Figure 1: Overview of the W-net approach. The first U-Net (Encoder) creates segmentations and the second U-Net (Decodedr) tries to reconstruct the image. During training both U-Nets get updated together to minimize the reconstruction loss. The segmented image has a few post processing steps, depicted as 'hierarchical segmentation' [2]

3.2 Deviation of the paper

The original code of the W-Net has not been shared and therefore the model had to be reproduced. The original paper misses out on some hyper-parameters which may have an effect on the behaviour of the W-Net. The paper misses the factor "k" for the k-way segmentation and the optimizer was not specified. The factor k will be varied in the experiments to decide which k fits the training data the best. For the optimizer Adam was chosen.

The paper states the ReLu non-linearity is applied before normalizing in each module. This study switched this order around and first applied batch normalization and then ReLu non-linearity. The paper suggested a learning schedule starting with a learning rate of 0.003 and divides it by 10 every 1.000 iterations. We found this made the network stop learning. A static learning rate of 0.003 was used. The post-processing step using hierarchical merging was not implemented, the conditional random field post-processing step has been implemented. Not adding the Hierarchical merging step resulted in more and smaller segments. This problem was partially covered during noise removal when mapping segments into bounding boxes, as described in section 3.4.

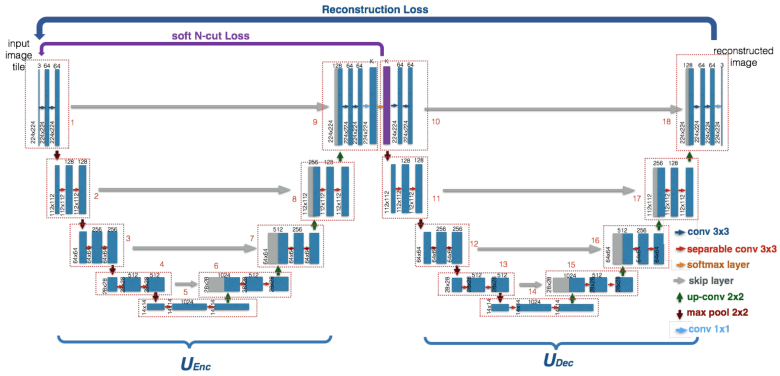


Figure 2: Overview of the auto-encoder. U_{enc} is the first U-Net encoder, the second U_{dec} is the which decodes the segmented image. [2]

3.3 Pre-processing images in the frequency domain

From the original training set two new training sets were created. On the first training set a custom filter was applied, on the second a high pass filter. The custom filter is based on the ground truth labelled training data. It tried to filter out the non-wheat head frequencies by creating a mask based on the difference between the frequencies inside- and outside the ground truth bounding boxes. Noise was removed from the mask by applying a square high-frequency filter of 120 pixels and a threshold of $0.5 * 10^{-7}$. The high pass filter is based on a simple mask of 40x40 pixels.

3.4 Create bounding boxes from semantic segments

To generate bounding boxes, a new post-processing step was introduced. This step takes the generated semantic segmentation from the W-Net and draws bounding boxes around each segmentation. Before drawing the bounding boxes noise was removed in three iterations using the morphological transformation Opening. After noise removal the bounding boxes were drawn based on the contours of the segments (Figure 3).

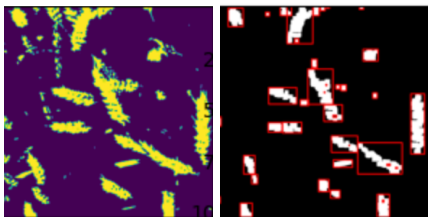


Figure 3: Creating bounding boxes around segments. Left: generated segmentation, right: drawn bounding boxes

3.5 Find the class representing wheat heads

The generated segmentations are classifying into k different classes. Since the W-Net is an unsupervised method, the classes are not labelled automatically. To find out which class represents the wheat heads the best, the accuracy was measured for each class. The class with the highest mean intersection over union (IoU) (1) was chosen to represent the wheat head class. Intersection over union is the overlap between the area of

two bounding boxes A and B divided by the total area of the two bounding boxes.

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \tag{1}$$

3.6 Evaluation

The performance of the W-Net has been tested using three cases:

1. Train and test the W-Net using wheat images without pre-processing
2. Train and test the W-Net using wheat images which are pre-processed using a custom filter
3. Train and test the W-Net using wheat images which are pre-processed using a high pass filter

To optimise the performance of the three cases, hyperparameter k was varied. The value k stands for the number of different classes the network can assign a segmentation between and has much influence on the result. The number k was varied between 16, 32 and 64. For the cases 1 and 2 k was also varied between 2, 3 and 4. In addition, for case 2 also $k = 128$ and $k = 256$ were chosen.

To measure the accuracy of the generated segmentation, the *precision value* and recall are calculated. The precision value (2) and recall (3) are based on the number of true and false positives (TP, FP) and false negatives (FN). They depend on a threshold t. The threshold t represents minimal IoU (1) score before its counted as a TP. The threshold is shown after the @ character. e.g. precision@0.01 means the precision measured if the threshold is $IoU > 0.01$.

$$Precision = \frac{TP(t)}{TP(t) + FP(t) + FN(t)} \tag{2}$$

$$Recall = \frac{tp}{tp + fn} \tag{3}$$

4 Experiments

The experiments are conducted using the Global Wheat Head Detection (GWHD) dataset. It contains 4700 images in total and has 190,000 human-annotated labelled wheat heads in the form of bounding boxes. The dataset consists of a large variety of genotypes and the images are collected from several countries across the world [11]. Since the W-Net model is unsupervised, the ground truth is not used during the training phase. The ground truth is only used to determine which class represents the wheat heads the best and to evaluate the data.

A subset of 500 images was randomly taken for GWHD dataset as the training set. The same images have been used to train all the models. The test set consists of 50 randomly chosen images from the GHWD dataset and do not overlap with the training set. During

training, the images were cropped to 224x224 pixels by always taking the top left most 224x224 pixels. The test data was not cropped. The architecture of the trained network is shown in Figure 2. A static learning rate of 0.003 was used, a dropout of 0.65 to prevent overfitting and the models were trained during 32 epoch with a batch size of 10. The models trained using k is 128 and 256 were trained with a batch size of 5 because of resource limitations.

4.1 Exp 1: Evaluate reproduced W-Net

To reproduce the W-Net, the repository from [1] was taken as the basis. Our W-Net generates more segmentations when compared to the original paper (Figure 4). This may be attributed to the fact that the W-Net model misses the hierarchical merging post-processing step. In Figure 5 the W-Net is compared to the papers segmentation without the hierarchical post-processing step and looked visually alike. Because the paper does not show the performance of the W-Net without the hierarchical grouping step, the comparison has only be done empirically. The goal of this research was not to create an exact reproduction of the W-Net but to research the influence of the frequency domain. The created W-Net was considered to be sufficient to answer the research question.

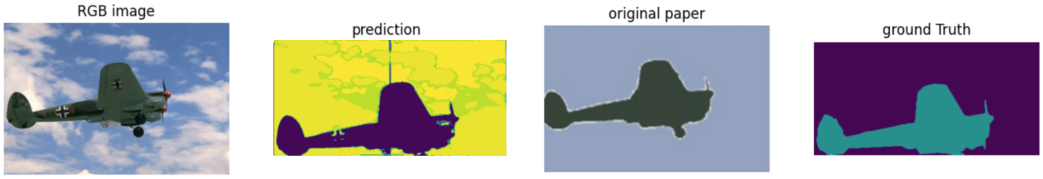


Figure 4: Segmented output of the W-Net. From left to right, the input image, the generated segmentation from our W-Net, the generated segmentation according to the paper, the ground truth segmentation. Our W-Net generates more segmentations compared to the paper.

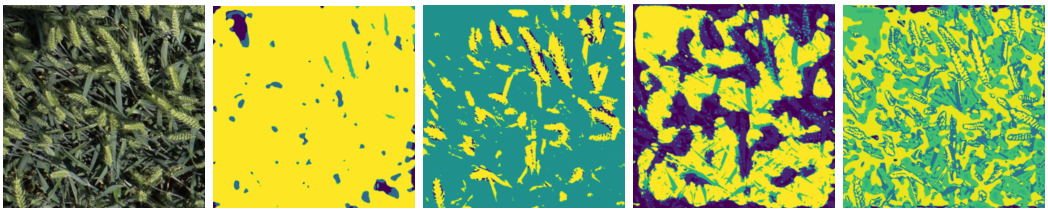
(a) (b)

Figure 5: When leaving out the hierarchical merging step the results look alike. a) Segmentation from our W-Net b) Segmentation without hierarchical merging according to the paper [2].

4.2 Exp 2: Investigate the accuracy without pre-processing

In this experiment the performance of the W-Net without pre-processing the training data was investigated. Six models have been trained with parameters k being 2, 3, 4, 16, 32 and 64. Setting k to 2 or 3, resulted in one single segment and therefore lost all its information. In Figure 6 the results of the segmentations are shown with the other k values. In Table 1

the mean IoU results are shown per variation. The model with parameter k=4 has the best mean IoU of 0.213.



(a) Original image (b) k=4 (c) k=16 (d) k=32 (e) k=64

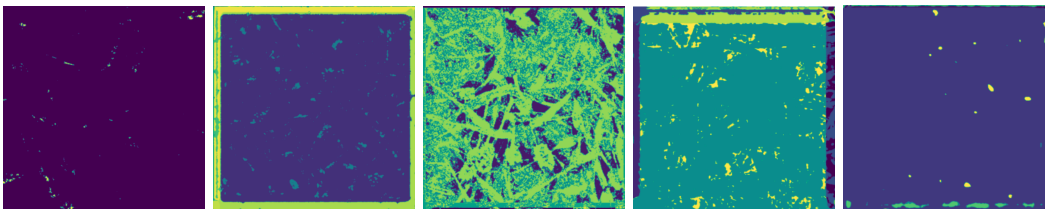
Figure 6: Segmentations generated without pre-processing the training data

model_nr	k	mean_iou	best_class	recall@0.01	precision@0.01	precision@0.5
31	2	0.173	1	0.0002	1	0.077
37	**4**	**0.213**	**2**	**0.384**	**0.663**	**0.161**
38	16	0.169	14	0.320	0.479	0.172
36	32	0.139	20	0.120	0.682	0.0556
32	64	0.144	40	0.186	0.616	0.099

Table 1: Performance of the models without pre-processing. Model 37 with $k = 4$ has the best mean IoU with 0.213.

4.3 Exp 3: Investigate the accuracy using a custom filter

In experiment 2 we have seen the accuracy of the W-Net without pre-processing the training data. Now the training data will be pre-processed using a custom filter as described in section 3.3. The W-Net has been trained with parameter k is 2, 3, 4, 16, 32, 64, 128 and 256. The choice to increase the number of k classes was motivated because the segmentations below $k = 64$ consistently converted into one segment. However, increasing the number to k is 128 or 256 did not improve the result. In Figure 7 the best segmented images are picked out, but as can be seen, also those do not segment well. In Table **??** the results of model 39 is shown. The mean IoU of the other models has not be calculated since they will be around zero.



(a) k=16 (b) k=32 (c) k=64 (d) k=128 (e) k=256

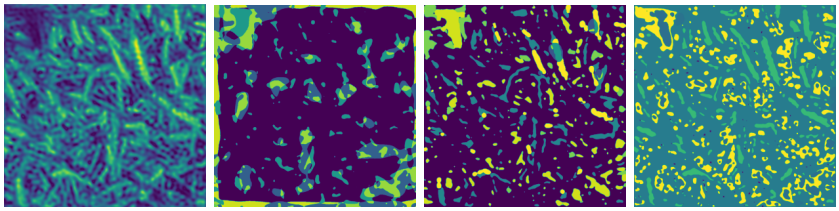
Figure 7: Segmentations generated when filtering was based on frequencies found on the bounding boxes inside the training data. Note, the best results were manually picked out of the test set. Most of the images segmented into one big segment.

model_nr	k	mean_iou	best_class	recall@0.01	precision@0.01	precision@0.5
43	4	0.036	0	0.190	1	0
41	16	0.178	16	0.190	0.0956	0
44	32	0.025	12	0.032	1	0
39	**64**	**0.110**	**12**	**0.037**	**0.857**	**0**

Table 2: The only result worth mentioning is the model with k=64. All other models failed to generate segmentations.

4.4 Exp 4: Investigate the accuracy using a high pass filter

Besides the custom filter the accuracy of the W-Net is tested on a high pass filter. The high pass filter uses a mask of 40x40 pixels. The W-Net has been trained with k is 16, 32 and 64. The segmentations showed promising results (Figure 8). The model with the best result as $k = 64$ with a mean IoU of 0.217 1. Based on the three trained models, it seems increasing the number of k classes, increases the mean IoU. The precision is 0.740 when a match is considered with an IoU above 0.01. The precision drops to 0.159 at an IoU threshold above 0.5 but does not drop as low as the other two models.



(a) Input image (b) k=16 (c) k=32 (d) k=64

Figure 8: Segmentations generated when pre-processing the training data with a high pass filter.

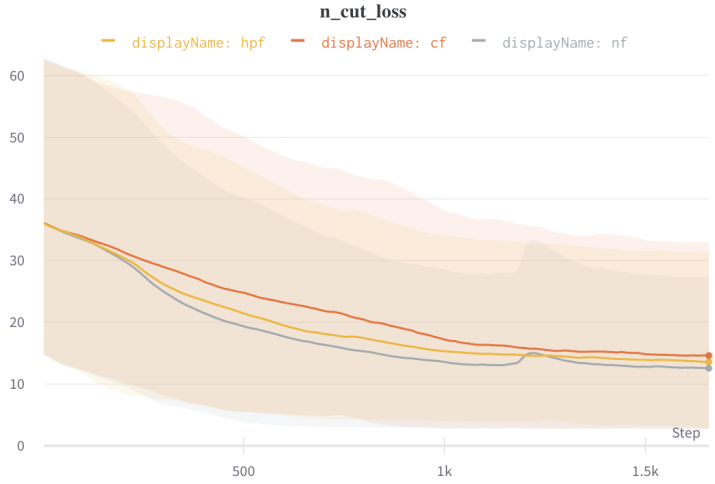
model_nr	k	mean_iou	class	recall@0.01	precison@0.01 @	precision@0.5
50	16	0.083	14	0.079	0.523	0.008
49	32	0.170	9	0.060	0.802	0.058
48	**0.217**	**38**	**38**	**0.150**	**0.740**	**0.159**

Table 3: Accuracy when the images are pre-processed using a high pass filter. Model 48, with k=64 has the best performance with a mean IoU of 0.216

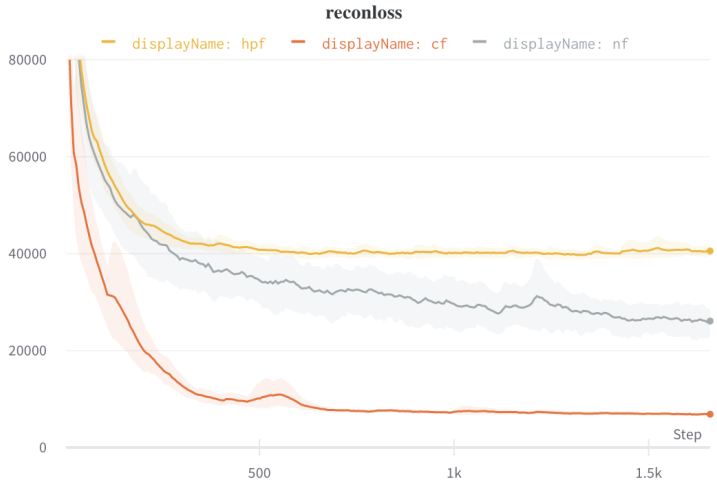
4.5 Comparison

With or without pre-processing the training data, the soft n-cut losses followed about the same pattern. Increasing the number of k classes increases the n-cut loss. At $k = 64$ the loss is around 30 and at $k = 16$ the loss is around 3.0 (Figure 9a). Pre-processing the data showed to create less deviation between the maximum and minimum reconstruction losses, compared to the models trained without pre-processed data (Figure 9b). The models trained with the custom filter showed a small reconstruction loss, but their accuracy the

worst. Looking at the deviation of the mean IoU, the deviation without pre-processing is 1,4 but the deviation when the high pass filter is applied is 8,7. The best models of the three cases are combined in Table 4. Pre-processing the images using the high-frequency filter increases the mean IoU by 1,49%.



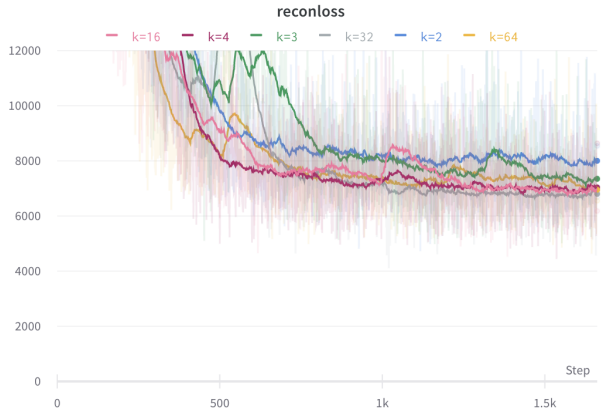
(a) N-cut losses, the different cases follow about the same pattern.



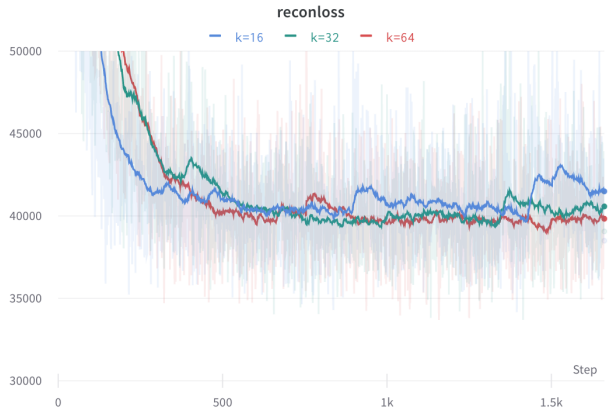
(b) Reconstruction loss. The deviation of the methods without pre-processing is larger then with pre-processing.

Figure 9: The mean losses of the models trained with k is 16, 32 and 64, grouped by pre-processing case. The padding around the lines is the difference between the maximum and minimum loss. np: models without pre-processing. cf, models trained with the custom filter data. hf: models trained with the high pass filter data.

(a) No pre-processing



(b) Custom filter



(c) High pass filter

Figure 10: Reconstruction loss of the different models. The deviation of the losses in a are larger compared to b and c *Graphs created using smoothing 0.95*

model_nr	k	mean_iou	filter	recall@0.01	precision@0.01	precision@0.5
37	4	0.213	none	0.384	0.663	0.163
39	64	0.110	custom filter	0.037	0.857	0
48	**64**	**0.217**	**high pass filter**	**0.150**	**0.740**	**0.159**

Table 4: Comparison between the best performing models. The model trained with the HPF gives the best mean IoU, but its precision is not the best.

5 Responsible Research

The field of computer vision can lead to many ethical concerns, especially in combination with facial recognition. This study does not raise such concerns directly, since we focus on the agricultural sector.

During this study the W-Net model needed to be reproduced. Some of its key hyper-parameters were undocumented and the code was not made public. This made reproducing the paper difficult and made us pay extra attention to the reproducibility of this paper.

To ensure reproducibility the setup and all its parameters are described in the paper, in addition the source code will be made available on https://gitlab.com/LifdAai/wheat-detection-wnet-frequency-information.git. All models trained for this study are numbered and their filename contains the hyperparameters. This makes the configuration of each model instantly visible. The results shown in the tables in section 4 contain model numbers that can be traced back in the code.

6 Discussion

We compared the precision of the W-Net with the baseline method associated with the GWHD dataset [11] in Table 5. The proposed approach failed to compete with the baseline method. The W-Net model generates semantic segmentation, but the ground truth labelling consists out of bounding boxes. The comparison, therefore, measures in two different measurements of truth and might be hard to compare directly. The method to convert segmentations into bounding boxes may have a significant impact on the accuracy of the W-Net.

model	precision@0.5
baseline	0.77
W-Net with HPF	0.16

Table 5: Comparison of the baseline method association with the GWHD dataset [11] and the best performing model, with the high pass filter.

6.1 Process

For this study the novel W-Net model has been used [2]. The authors did not share their code and some details in their paper were missing to fully be able to reproduce their paper.

To use the W-Net more than 12 reproductions have been considered, but none of them had managed to reproduce the results of the original paper. From the 12 reproductions a top 3 was chosen and tested and the best performing reproduction had been chosen for this study. After choosing the repository and trying to get some statistics, an extra 'post-processing' that did not follow the original paper was found. This step tried to map the unsupervised segmentations with the actual ground truth segmentations. This step optimized the results and should not have been in the reproduction. The whole code had to be scanned from top to bottom and all ambiguities had to be removed.

The work put into finding a repository, understanding, debugging and restoring the ambiguities took considerable time. Documenting this process was not the topic of this study and therefore not included in this paper but had an effect on the final result. We would like to dive deeper in the analysis of the precision.

7 Conclusions and Future Work

In this study we have tried to answer the following question:

"How can frequency information be used to improve the accuracy of the unsupervised segmentation model W-Net, when applied to identify wheat heads in images."

Three datasets were created. The first dataset was a subset of the Global Wheat Head Detection dataset. The other two sets consist of the same set of images but modified in the frequency domain. One set was modified using a custom filter (CF), the other was modified using a high pass filter (HPF). The custom filter tried to filter all other frequencies unique to the ground truth training data.

The segmentations outputted by the W-Net were mapped to bounding boxes based on the contours of the segments. The training set without modifications had an intersection over union (IoU) of 0.213, using a high pass filter this was slightly improved with 1,4% to 0.217. The custom filter had an IoU of only 0.110. The significance of the 1,4% might be debatable because the training set (500 images) and test set (50 images) are relatively small compared to the 4700 images in the GWHD dataset and can easily be increased in size multiple times.

An interesting finding is that using frequency information seems to affect the deviation of the reconstruction loss (standard deviation of 33 (CF) and 858 (HPF) against 2514). This might be caused because the frequency filters might have filtered out the noise.

Another interesting finding was that the reconstruction loss of models trained using the custom filter was significantly lower compared to the other models, but did not result in a high IoU score. The low reconstruction loss could be caused by the dark blue monotonic images produced by the custom filter. Normalizing the colours of these images and retrain the model might be investigated in further research.

Future work can be done by evaluating the W-Net wheat head detection on a dataset consisting of segmentations as ground truth instead of bounding boxes.

The precision with threshold IoU 0.01 shows high precision. This may indicate that the location of the segmentation is correct, but the size is incorrect. A future study could

analyse what percentage of the generated segmentation lies inside the ground truth, instead of taking the IoU.

8 Acknowledgement

I am indebted to Attila Lengyel, Nergis Tomen, Yancong Lin and Silvia Pintea for their advice and supervision during the research and would like to thank Elvin Isulfi for examining the research paper. I would also thank my fellow students Dani Rogmans, Alin Prundeanu, Nick Mertzanis and Petar Ulev for their collaboration during the project.

References

[1] "Nadine Duursma" "Guru Deep Singh". *W-Net: A Deep Model for Fully Unsupervised Image Segmentation Reproduction*. 2021. URL: https://gurudeep1998.medium.com/w-net-a-deep-model-for-fully-unsupervised-image-segmentation-reproduction-2651540eaed6.

[2] "Brian Kulis" "Xide Xia". "W-Net: A Deep Model for Fully Unsupervised Image Segmentation". In: *CoRR* abs/1711.08506 (2017). arXiv: 1711.08506. URL: http://arxiv.org/abs/1711.08506.

[3] "Xu Z" "Zhou C" "Hu J". "A Monitoring System for the Segmentation and Grading of Broccoli Head Based on Deep Learning and Neural Networks". In: *Frony. Plant Sci.* 402.11 (2020). DOI: https://doi.org/10.3389/fpls.2020.00402.

[4] P ArbelÃ¡ez et al. "Contour Detection and Hierarchical Image Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.5 (May 2011), 898â916. DOI: 10.1109/tpami.2010.161.

[5] Pablo Arbelaez et al. "Contour Detection and Hierarchical Image Segmentation". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 33.5 (May 2011), pp. 898–916. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2010.161. URL: http://dx.doi.org/10.1109/TPAMI.2010.161.

[6] Pablo Arbelaez et al. *From contours to regions: An empirical evaluation*. June 2009. DOI: 10.1109/CVPR.2009.5206707. URL: https://ieeexplore.ieee.org/document/5206707.

[7] Muhammad Hamza Asad and Abdul Bais. "Weed detection in canola fields using maximum likelihood classification and deep convolutional neural network". In: *Information Processing in Agriculture* 7.4 (2020), pp. 535–545. ISSN: 2214-3173. DOI: https://doi.org/10.1016/j.inpa.2019.12.002. URL: https://www.sciencedirect.com/science/article/pii/S2214317319302355.

[8] Wenlin Chen et al. "Compressing Convolutional Neural Networks in the Frequency Domain". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 22 (Aug. 2016). DOI: 10.1145/2939672.2939839.

[9] Dorin Comaniciu and Peter Meer. "Mean shift: A robust approach toward feature space analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.5 (2002), pp. 603–619.

[10] Timothée Cour, Florence Bénézit, and J. Shi. "Spectral segmentation with multiscale graph decomposition". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* 2 (2005), 1124–1131 vol. 2.

[11] E. David 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*. 2020. arXiv: 2005.02162 [cs.CV].

[12] "Deep learning and machine vision for food processing: A survey". In: *Current Research in Food Science* 4 (2021), pp. 233–249. ISSN: 2665-9271. DOI: https://doi.org/10.1016/j.crfs.2021.03.009. URL: https://www.sciencedirect.com/science/article/pii/S2665927121000228.

[13] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. "Efficient Graph-Based Image Segmentation". In: *International Journal of Computer Vision* 59.2 (Sept. 2004), 167â181. DOI: 10.1023/b:visi.0000022288.19776.77.

[14] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. "Speeding up Convolutional Neural Networks with Low Rank Expansions". In: *arXiv:1405.3866 [cs]* (May 2014). URL: https://arxiv.org/abs/1405.3866v1.

[15] Xue Jiao, Yonggang Chen, and Rui Dong. "An unsupervised image segmentation method combining graph clustering and high-level feature representation". In: *Neurocomputing* 409 (2020), pp. 83–92. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2020.05.073. URL: https://www.sciencedirect.com/science/article/pii/S0925231220309243.

[16] D. Martin et al. "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics". In: *Proc. 8th Int'l Conf. Computer Vision*. Vol. 2. July 2001, pp. 416–423.

[17] George.E Meyer et al. "Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images". In: *Computers and Electronics in Agriculture* 42.3 (2004), pp. 161–180. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2003.08.002. URL: https://www.sciencedirect.com/science/article/pii/S0168169903001224.

[18] Shervin Minaee et al. "Image Segmentation Using Deep Learning: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1. DOI: 10.1109/TPAMI.2021.3059968.

[19] Varsha Nair et al. "Fast Fourier Transformation for Optimizing Convolutional Neural Networks in Object Recognition". In: *arXiv preprint arXiv:2010.04257* (2020).

[20] Varsha Nair et al. "Fast Fourier Transformation for Optimizing Convolutional Neural Networks in Object Recognition". In: *arXiv:2010.04257 [cs, eess]* (Oct. 2020). URL: https://arxiv.org/abs/2010.04257.

[21] Thanh Minh Nguyen and QM Jonathan Wu. "Fast and robust spatially constrained Gaussian mixture model for image segmentation". In: *IEEE transactions on circuits and systems for video technology* 23.4 (2012), pp. 621–635.

[22] Navid Nikbakhsh, Yasser Baleghi, and Hamzeh Agahi. "A novel approach for unsupervised image segmentation fusion of plant leaves based on G-mutual information". In: *Machine Vision and Applications* 32 (Oct. 2020), p. 5. DOI: 10.1007/s00138-020-01130-0. URL: https://link.springer.com/article/10.1007%2Fs00138-020-01130-0#citeas (visited on 06/25/2021).

[23] C. Fowlkes P. ArbelÃ¡ez M. Maire and J. Malik. "Contour Detection and Hierarchical Image Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.5 (May 2011), pp. 898–916. ISSN: 0162-8828.

[24] Lihong Xu Ping Zhang. "Unsupervised Segmentation of Greenhouse Plant Images Based on Statistical Method". In: *Nature* 8.4465 (2018). DOI: https://doi.org/10.1038/s41598-018-22568-3. URL: https://www.nature.com/articles/s41598-018-22568-3.

[25] Silvia Pintea. *Deep learning for precision agriculture*. Delft University of Technology. URL: https://projectforum.tudelft.nl/course_editions/39/projects/980.

[26] Olaf Ronneberger, Philaipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[27] University of Saskatchewan. URL: https://www.kaggle.com/c/global-wheat-detection/overview.

[28] L. Senni et al. "On-line automatic detection of foreign bodies in biscuits by infrared thermography and image processing". In: *Journal of Food Engineering* 128 (2014), pp. 146–156. ISSN: 0260-8774. DOI: https://doi.org/10.1016/j.jfoodeng.2013.12.016. URL: https://www.sciencedirect.com/science/article/pii/S0260877413006262.

[29] Amos Sironi et al. "Learning Separable Filters". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.1 (Jan. 2015), 94â106. DOI: 10.1109/tpami.2014.2343229.

[30] Artem Vasilyev. "CNN optimizations for embedded systems and FFT". In: *Standford University Report* (2015).

[31] Rosemarie Velik. "Discrete fourier transform computation using neural networks". In: *2008 International Conference on Computational Intelligence and Security*. Vol. 1. IEEE. 2008, pp. 120–123.

[32] Kyosuke Yamamoto et al. "On Plant Detection of Intact Tomato Fruits Using Image Analysis and Machine Learning Methods". In: *Sensors* 14.7 (2014), pp. 12191–12206. ISSN: 1424-8220. URL: https://www.mdpi.com/1424-8220/14/7/12191.