Apple Counting Network Before Fruit Thinning Period Based On Dilated Convolution

Mengyang Song
School of Software Henan
Polytechnic University, Jiaozuo
454000, China

Guoquan Jiang
School of Software Henan
Polytechnic University, Jiaozuo
454000, China

Zhanqiang Huo*
School of Software Henan
Polytechnic University, Jiaozuo
454000, China

Zhengyuan Yang
School of Software Henan
Polytechnic University, Jiaozuo
454000, China

Hongxu Zhang
School of Software Henan
Polytechnic University, Jiaozuo
454000, China

DONE

ABSTRACT

Abstract: Fruit counting is an integral part of achieving precision orchard management. Accurate counting of the number of fruits on a tree can provide critical information for yield estimation, thus promoting precision agriculture. However, today fruit farmers can only support their fieldwork by manual counting, and a reliable and accurate automatic fruit counting method is missing. A pre-thinning apple counting network (FTACNet) is proposed to address the problems of shading, uneven distribution, and fruit scale differences and is validated on the produced dataset. The method uses deep learning algorithms in the field of population counting. FTACNet shows good performance on the dataset, with mean absolute error (MAE) down to 4.14 and mean square error (MSE) down to 5.62. Moreover, the model is end-to-end, the model is small, and can be easily deployed to mobile devices, which has good potential for application in orchards.

Ref. against what

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Computer vision; Computer vision problems; Object detection; Machine learning; Machine learning approaches; Neural networks; • **Mathematics of computing** → Probability and statistics; Probabilistic inference problems; Density estimation.

KEYWORDS

Key words: Crowd counting, Deep learning, Fruit counting, Precision agriculture

ACM Reference Format:
Mengyang Song, Guoquan Jiang, Zhanqiang Huo, Zhengyuan Yang, and Hongxu Zhang. 2022. Apple Counting Network Before Fruit Thinning Period Based On Dilated Convolution. In *2022 The 11th International Conference on Networks, Communication and Computing (ICNCC) (ICNCC 2022), December 09–11, 2022, Beijing, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3579895.3579915

1 INTRODUCTION

1

Precision agriculture [1] is agricultural management's observation, measurement, and response to crop growth to optimize the return on inputs while conserving resources. The development of precision agriculture will make crop or livestock farming more precise and controllable.

Precision agriculture provides excellent technical support for accurate yield statistics. Accurate yield statistics are essential for agricultural development, especially for fruit, which has recently been a hot topic. Apples are one of the crops grown on a large scale in China and are an integral part of the agricultural economy.

According to the UN FAO statistical database, the world area under apple cultivation reached 69.94 million mu in 2020, with an annual world production of 89.73 million tons. Therefore, orchard production management to increase production in the apple industry becomes particularly important.

Obtaining information on the number of cash crops such as apples is essential for orchard production management by fruit farmers. Apples typically drop flowers and fruits four times during the growing process, from bud emergence to fruit harvest [2]. Before fruit picking, large changes in temperature, soil dryness, and humidity disorders, as well as a large amount of fruit tree fruiting, can lead to increased fruit drop, thus causing greater economic losses to the fruit farmers. Therefore, during this period, it is necessary to evaluate the yield of the orchard several times in order to adjust management measures in time to reduce the number of fruit drops.

Accurate calculation of the number of fruits on the tree can provide critical information for yield estimation and support effective precision orchard management. However, people can only support their fieldwork by manual counting, and a reliable and accurate automatic fruit counting method is missing. Since it is a costly and labor-intensive task to count all the fruits in an orchard by manual counting alone, the development of a reliable and accurate automatic fruit-counting method has been a strong demand from fruit farmers.

3

Although some machine learning methods [3] have been implemented on related tasks such as yield estimation in the last few

[*] The corresponding author: hzq@hpu.edu.cn

ICNCC 2022, December 09–11, 2022, Beijing, China

ACM ISBN 978-1-4503-9803-9/22/12…$15.00
https://doi.org/10.1145/3579895.3579915

years, the introduction of deep learning in recent years has had a significant impact on the field.

CNN The most popular neural network currently is the deep convolutional neural network, and convolutional networks are currently achieving great success in many research areas, such as fruit counting [4], crowd counting, image segmentation [5], natural language processing, etc. 1. 2.

6 Most fruit counting methods are detection followed by counting, and their essential task is still detection rather than counting. * Among the population counts, the favored method nowadays is based on density maps [6], which has the advantage of being less sensitive to occlusion and is more robust compared to the method of detecting first and counting later.

CSR NET In this paper, we introduce FTACNet, a network for counting apples before the thinning period, to address the problem of time-consuming and labor-intensive manual counting in modern apple orchards. We demonstrate that using common tools such as smartphones and combining deep learning networks used for object and crowd counting will work well for counting apples in orchards.

For model performance metrics, we use mean absolute error (MAE) and mean square error (MSE), which are the most commonly used evaluation metrics for agricultural yield estimation with population counts.

2 RELATED WORK

The prevailing method of fruit counting is by first detecting the fruit and then directly counting the detected fruit, so it is still a fruit detection study.

Bargoti and Underwood [7] used WS and circular Hough transform (CHT) algorithms for pixel-level segmentation and processing of target images to detect and count individual fruits. Dorj et al [8] studied the detection and counting of citrus by watershed segmentation (WS) algorithm based on the feature of the color of the fruit. Mekhalfi et al. [9] developed a sensor system based on the Viola-Jones target detection algorithm to detect kiwi fruits in the field and count them. Stein et al. [10] first photographed from multiple angles of the fruit tree, then performed 3D reconstruction of the fruit, and then used the FasterR-CNN network for fruit detection and counting. Convolutional (CNN) architectures and several of their variants (e.g., MaskR-CNN [11]) have become classical for yield estimation [12, 13]. 11 12 13

The fruit counting problem can be combined with methods in the field of population counting, and the current state-of-the-art in this field [14, 15] treats population counting as a density map estimation problem. For images of large populations, this density map estimation method has been shown to be more robust than the detection and then counting method because the former is less sensitive to occlusion and it does not require binarization decisions at an early stage [16]. 16

17 In 2018, Li et al. [17] proposed a dilated convolutional neural network model CSRNet for dense crowd counting, which is a single-channel counting network, and thus, the number of parameters in the network is significantly reduced, and the difficulty of training is reduced. The success of CSRNet provided new ideas in the field of crowd counting, and subsequently, many scholars started to follow the research in the field of crowd counting using dilated convolution [18]. 18 Although crowd-counting solutions based on deep neural networks have shown good results, the effectiveness of the technique is still severely affected by factors such as occlusion and differences in crowd distribution in highly crowded and complex scenes. To this end, Liu et al. [19] proposed a deformable convolutional network ADCrowdNet incorporating attention mechanisms for crowd counting, and better counting results were obtained.

Domestic and foreign researchers have provided the oretical basis and technical support for apple recognition and counting based on the application of deep learning in fruit detection [20, 21], as well as deep learning methods used in other fields. However, at present, due to the influence of different lighting, background, and other environments with human factors such as shooting height and shooting distance, it poses a greater challenge to the accuracy and stability of deep learning-based apple counting.

Here, for the demand of apple counting in orchards of the majority of orchard farmers, we propose FTACNet for the problems of occlusion, fruit scale differences, uneven distribution of fruits, etc. The method uses methods in deep learning to learn the features of different apples, extract some implicit information about apples in pictures to count the fruits, and compare and analyze with other counting methods.

We combined the recent deep neural networks for crowd and object counting [22, 23] to provide critical information for orchard yield estimation by automatically estimating the number of apples for effective orchard precision management [24].

3 PROPOSED SOLUTION

3.1 Dataset Production

This experiment was conducted on apple trees with fruit. The fruit from apple trees is usually spherical, and the mature apple is usually red. We collected pictures before the thinning period of apple growth, mostly lime green, and the fruit usually grows in aggregates. The apples are relatively similar in color to the leaves, which are not easily distinguishable.

Since apples and kiwis, pears, peaches, and other fruits are similar in appearance and have a wide representation of fruits, the research results based on apples can be extended to other kinds of fruits, and the research has good development significance.

FTACNet is based on the use of smartphone images and deep learning methods to estimate the number of apples in an image. In this study, images were collected using smartphones, and data were collected before the apple thinning period. The data was collected under as many lighting conditions as possible, including downlighting, metering, and backlighting. Images were taken and selected using a smartphone, resulting in 862 images forming the dataset for training and testing the deep learning network.

While previous fruit counts were labeled using a rectangular box to frame the target, we used a different method when labeling images. * Since the network was developed in the context of automatic crowd counting, we use a common labeling method in the field of crowd counting, i.e., point labeling, where a point is labeled in the center of the apple. Annotate all the captured images one by one, i.e., annotate the center of the apple in the image. Among them, only apples that can be manually identified or the outline of the

MANUAL LABELLING

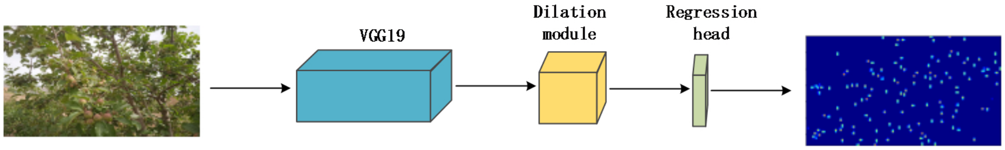

Figure 1: Overall structure diagram

fruit can be inferred are labeled, and fruits that fall on the ground are not labeled.

The labeled completed images are used as the data set, where the training set has 689 images and the corresponding labeled files, and the test set has 173 images and the corresponding files.

3.2 From Crowd Counting To Apple Counting

FTACNet uses dilated convolution and integrates a Bayesian loss function [25] to handle the apple counting problem. The capabilities of FTACNet are then demonstrated on our labeled dataset. Compared with the original Bayesian loss function used in the network and other counting methods, FTACNet has a better performance in the mean absolute error and mean square error (MAE/MSE) metrics. We can well illustrate that for a task like fruit counting, some methods developed in the field of population counting can be used. An image is passed into our FTACNet model, which is supervised using a Bayesian loss function, and the predicted density map is then output, and the number of apples is calculated by integrating the density map. The whole process is shown in Figure 1.

CSRNET RELEVANT

The labeling in crowd counting is to label each head, then applied to apple counting is to label each apple. The commonly used methods are converting the annotation to a density map after the Gaussian kernel as GT, but this GT is not the most suitable considering the apples-to-apples occlusion, the leaves-to-apples occlusion, and the view angle, etc. When tested on real ground truth data, the use of a Gaussian kernel to smooth the corresponding point annotations can cause damage to the generalization range of the model [16]. We thus use a Bayesian loss function as a point-supervised estimate of the number of apples. The original Gaussian kernel function transforms the point labeling into a ground truth density map, and the loss function we use provides a more appropriate supervision of the count expectation for each labeled point. The loss function is used to supervise the training model to generate better prediction density maps.

3.3 Network Structure

3.3.1 Dilated Convolution. A key part of the network structure we designed is the dilated convolution layer. 2-D dilated convolution [26] can be defined as following:

$$y\,(m,n)=\sum_{i=1}^{M}\sum_{j=1}^{N} x\,(m+r\times i, n+r\times j)\; w\,(i,j)$$

$y(m,n)$ is the result of the dilated convolution of a filter $w(i,j)$ with length and width M and N, respectively, with the input. $x(m,n)$ Where the parameter r represents the dilated rate of the dilated convolution. When r =1, this dilated convolution becomes a normal convolution.

Dilated convolution is widely used in tasks such as semantic segmentation and target detection. Dilated convolution is roughly the same as a standard convolution in terms of implementation, and standard convolution can be understood as a special form of dilated convolution. In various tasks, downsampling is needed to increase the network's field of perception while reducing computational effort, which increases the field of perception but decreases the resolution at the same time. To be able to expand the perceptual field while preserving the resolution, one can use dilated convolution. The property of dilated convolution expands the perceptual field without adding an additional number of parameters or computational effort [17].

3.3.2 Network Configuration. The front end of our FTACNet model is VGG19. And the back end, which is our dilated convolution module. Network configuration is shown in Figure 1, where the dilated convolution can be done with different dilation rates, and the most appropriate dilation rate needs to be verified by experiments. The detailed analysis is shown in Section 4.2 of the article.

27

The FTACNet model uses the standard image classification network VGG19 [27] as the first module, which is pre-trained on ImageNet [28] and removes the last pool of this network as well as the fully connected layers that follow it. This module is used as our feature extractor. After the input image is output from this module, it is then passed to our added dilated convolution module, the structure of which is shown in Figure 2, where the parameter meaning of the dilated convolution layer is "convolution - (kernel size) - (number of filters) - (dilation rate)." The advantage of using dilated convolution is that the perceptual field is expanded without pooling to lose information, thus allowing each convolution output to contain a larger range of information. At the same time, we do not introduce additional parameters that would increase the computational effort.

We use bilinear interpolation to upsample the output of the first module, sampling to 1/8 of the input image size, and then pass it through the six-layer dilated convolution we added, followed by passing in the regression head, which is a 1x1 convolutional layer, to finally obtain our predicted density map.

3.3.3 Training Details . We use horizontal flipping and random cropping to expand the data used for training. For random cropping for training, we set the crop size to 512x512. We use a Bayesian loss function to evaluate the difference between our generated estimated density maps and the ground truth. An ADAM optimizer with an initial learning rate of 10-5 is used to update the parameters.

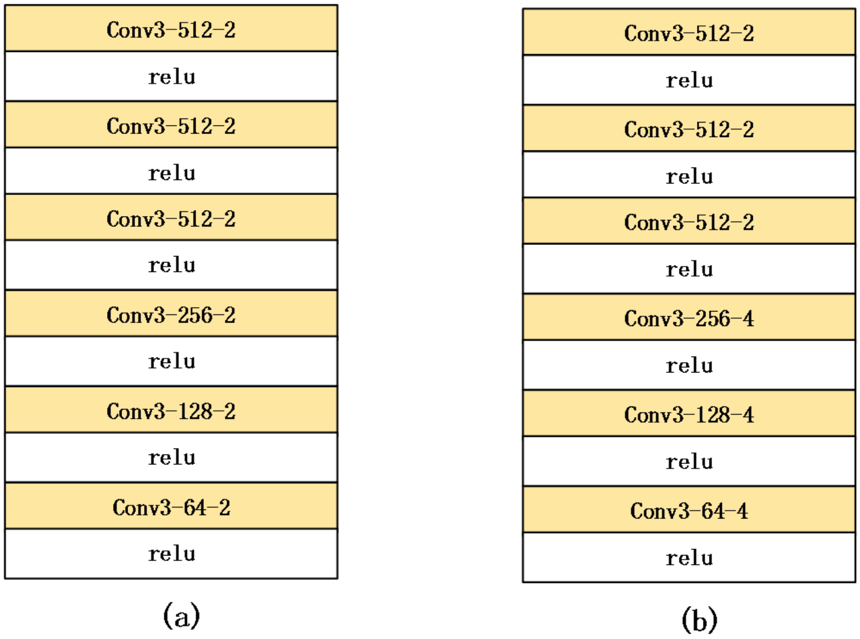

Figure 2: (a) dilation module in FTACNet A. (b) dilation module in FTACNet B

Table 1: Comparing the results of different dilation rates

Architecture MAE	MSE
FTACNet A 4.70	6.08
FTACNet B 4.14	5.62

4 EXPERIMENTS

4.1 Evaluation Metrics

To evaluate the performance of FTACNet models, we use the most commonly used metrics in the field of agricultural yield estimation and population counting, namely mean absolute error (MAE) and mean square error (MSE). These two metrics are defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_i - C_i^{GT}\right|$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_i - C_i^{GT}\right)^2}$$

where C_i is the estimated count and C_i^{GT} is the ground truth count corresponding to image i. The estimated counts are the result of the integration of the output density map. These two metrics are a measure of the accuracy (MAE) and robustness (MSE) of the model.

4.2 Ablation Experiments

We conducted ablation experiments to analyze the effect of different dilation rates on FTACNet using our apple dataset. FTACNet A is a six-layer dilated convolution with a dilation rate of 2 for all six layers, and FTACNet B is a combination of 2 and 4 with a dilation rate of 2 for the first three layers and 4 for the last three layers of the six-layer dilated convolution.

The experimental results are shown in Table 1. After comparison, FTACNet B has the best results, so we use FTACNet B as the proposed FTACNet for the following experiments.

4.3 Prediction Results Show

The predicted density map of our proposed FTACNet model is shown in Figure 3, and the number of apples calculated from the predicted density map is very close to the actual value.

4.4 Evaluation And Comparison

The experimental results are shown in Table 2, where CSRNet+ is the method we adapted to generate density maps in CSRNet to fix the kernel size and make it equal to 3. Such a method of generating density maps is more suitable for our dataset. Table 2 compares FTACNet with various other methods. In all experiments, our method outperforms several other methods, and the MAE of FTACNet is reduced by 3.5% and the MSE is reduced by 5.4% compared to the Bayesian method. The experimental results show that the performance of our proposed FTACNet method is good.

The experimental results of our proposed method and the Bayesian loss function method [25] are compared, and the results

Table 2: Results of the various methods on our dataset

Method	Backbone	MAE	MSE
CSRNet [17]	VGG16	9.79	12.71
CSRNet+	VGG16	5.85	7.52
Bayesian Loss [25]	VGG19	4.29	5.94
FTACNet(proposed)	VGG19	4.14	5.62

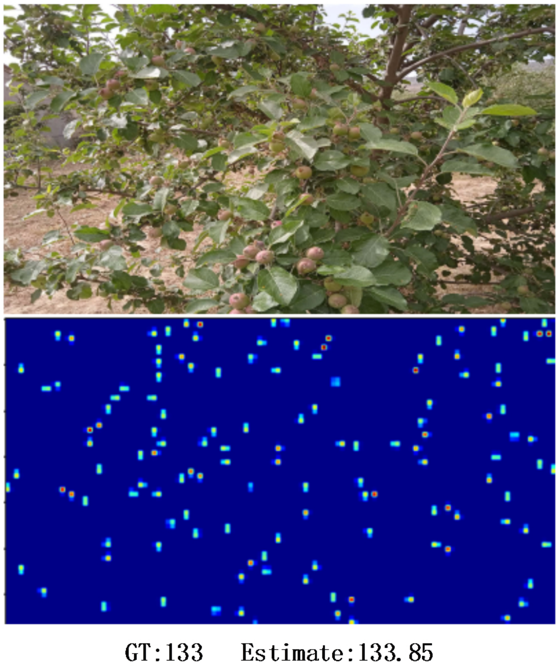

Figure 3: density map predicted by FTACNet

are shown in Figure 4. The first row of images is the original data, the second row of images is the density map prediction result of the BayesianLoss method, and the last row is our method's density map prediction result. Under each image is the corresponding Ground Truth or the predicted number of apples. The comparison shows that the number of fruits obtained by the algorithm in this paper is more similar to the actual value.

5 CONCLUSION

This paper demonstrates that automatic apple counting can be performed well-using tools such as smartphones and through the FTACNet model. FTACNet Method Enables Accurate Fruit Counting. In this study, the counting was performed on apples, which are more similar in shape to fruits such as green oranges and kiwis and are more representative, so the research results can be extended to many other fruits and have good research significance. In addition, the model is end-to-end and has good application potential in orchards as the model is small and can be easily deployed to mobile devices.

The data in the current dataset were collected before the fruit thinning stage. The next step can be considered to collect data at multiple growth stages of apples, such as the flowering and ripening stages, and then train them so that the model can be used at multiple growth stages of apples to increase the robustness of the model.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (62273292) and Henan University Scientific and Technological Innovation Team Support Program 19IRTSTHN012).

REFERENCES

[1] Yang, X., Shu, L., Chen, J., Ferrag, M. A., Wu, J., Nurellari, E. and Huang, K. A survey on smart agriculture: Development modes, technologies, and security and privacy challenges. IEEE/CAA Journal of Automatica Sinica, 8, 2 (2021), 273-302.

[2] Genno, H. and Kobayashi, K. Apple growth evaluated automatically with high-definition field monitoring images. Computers and Electronics in Agriculture, 164 (2019), 104895.

[3] Sei, Y. and Ohsuga, A. Count Estimation With a Low-Accuracy Machine Learning Model. IEEE Internet of Things Journal, 8, 8 (2020), 7079-7088.

[4] Gao, F., Fang, W., Sun, X., Wu, Z., Zhao, G., Li, G., Li, R., Fu, L. and Zhang, Q. A novel apple fruit detection and counting methodology based on deep learning and trunk tracking in modern orchard. Computers and Electronics in Agriculture, 197 (2022), 107000.

[5] Peng, B., Zhang, L., Mou, X. and Yang, M.-H. Evaluation of segmentation quality via adaptive composition of reference segmentations. IEEE transactions on pattern analysis and machine intelligence, 39, 10 (2016), 1929-1941.

[6] Zhang, C., Li, H., Wang, X. and Yang, X. Cross-scene crowd counting via deep convolutional neural networks. City, 2015.

[7] Bargoti, S. and Underwood, J. P. Image segmentation for fruit detection and yield estimation in apple orchards. Journal of Field Robotics, 34, 6 (2017), 1039-1060.

[8] Dorj, U.-O., Lee, M. and Yun, S.-s. An yield estimation in citrus orchards via fruit detection and counting using image processing. Computers and Electronics in Agriculture, 140 (2017), 103-112.

[9] Mekhalfi, M. L., Nicolò, C., Ianniello, I., Calamita, F., Goller, R., Barazzuol, M. and Melgani, F. Vision system for automatic on-tree kiwifruit counting and yield estimation. Sensors, 20, 15 (2020), 4214.

[10] Stein, M., Bargoti, S. and Underwood, J. Image based mango fruit detection, localisation and yield estimation using multiple view geometry. Sensors, 16, 11 (2016), 1915.

[11] He, K., Gkioxari, G., Dollár, P. and Girshick, R. Mask r-cnn. City, 2017.

[12] Silver, D. L. and Monga, T. In vino veritas: Estimating vineyard grape yield from images using deep learning. Springer, City, 2019.

[13] Škrabánek, P. DeepGrapes: Precise Detection of Grapes in Low-resolution Images. IFAC-PapersOnLine, 51, 6 (2018), 185-189.

[14] Xu, C., Qiu, K., Fu, J., Bai, S., Xu, Y. and Bai, X. Learn to scale: Generating multipolar normalized density maps for crowd counting. City, 2019.

[15] Zhang, Q. and Chan, A. B. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. City, 2019.

[16] Wang, B., Liu, H., Samaras, D. and Nguyen, M. H. Distribution matching for crowd counting. Advances in neural information processing systems, 33 (2020), 1595-1607.

[17] Li, Y., Zhang, X. and Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. City, 2018.

[18] Liu, W., Salzmann, M. and Fua, P. Context-aware crowd counting. City, 2019.

[19] Liu, N., Long, Y., Zou, C., Niu, Q., Pan, L. and Wu, H. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. City, 2019.

[20] Aggelopoulou, A., Bochtis, D., Fountas, S., Swain, K. C., Gemtos, T. and Nanos, G. Yield prediction in apple orchards based on image processing. Precision Agriculture, 12, 3 (2011), 448-456.

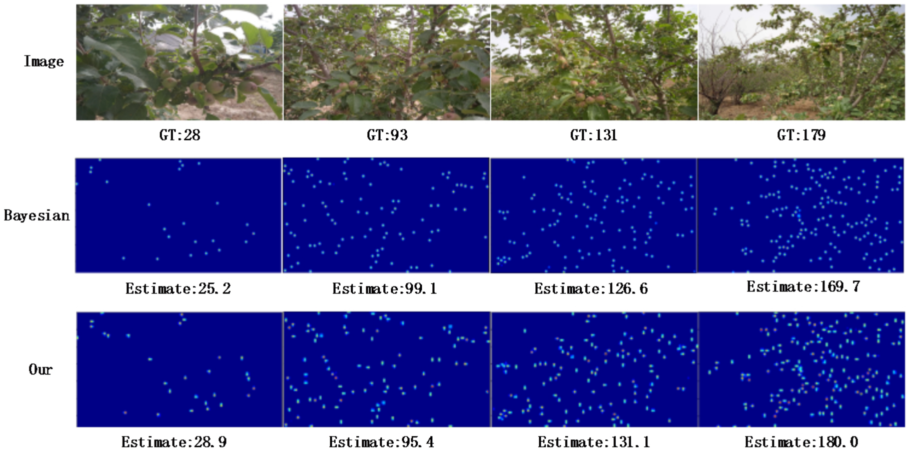

Figure 4: comparison of the results of our method and Bayesian Loss method

[21] Ukwuoma, C. C., Zhiguang, Q., Bin Heyat, M. B., Ali, L., Almaspoor, Z. and Monday, H. N. Recent advancements in fruit detection and classification using deep learning techniques. Mathematical Problems in Engineering, 2022 (2022).

[22] Ren, S., He, K., Girshick, R. and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28 (2015).

[23] Xie, W., Noble, J. A. and Zisserman, A. Microscopy cell counting and detection with fully convolutional regression networks. Computer methods in biomechanics and biomedical engineering: Imaging & Visualization, 6, 3 (2018), 283-292.

[24] Coviello, L., Cristoforetti, M., Jurman, G. and Furlanello, C. GBCNet: In-field grape berries counting for yield estimation by dilated CNNs. Applied Sciences, 10, 14 (2020), 4870.

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

[26] Yu, F. and Koltun, V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015).

[27] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[28] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. Ieee, City, 2009.