AN AI-HORTICULTURE MONITORING AND PREDICTION SYSTEM WITH AUTOMATIC OBJECT COUNTING

A Thesis

by

XUETING LIU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Xiaoning Qian
Committee Members,	Byung-Jun Yoon
	I-Hong Hou
	Zhangyang Wang
Head of Department,	Miroslav M. Begovic

December 2019

Major Subject: Electrical Engineering

ABSTRACT

Estimating density maps and counting the number of objects of interest from images has a wide range of applications, such as crowd counting, traffic monitoring, cell microscopy in biomedical imaging, plant counting in agronomy, as well as environmental survey. Manual counting is a labor-intensive and time-consuming process. Over the past few years, the topic of automatic object counting by computers has been actively evolving from the classic machine learning methods based on handcrafted image features to end-to-end deep learning methods using data-driven feature engineering, for example by Convolutional Neural Networks (CNNs).

In our research, we focus on the task of counting plants for large-scale nursery farms to build an AI-horticulture monitoring and prediction system using unmanned aerial vehicle (UAV) images. The common challenges of automatic object counting as other computer vision tasks are scenario difference, object occlusion, scale variation of views, non-uniform distribution, and perspective difference. For an AI-horticulture monitoring and prediction system for large-scale analysis, the plant species various a lot, so that the image features are different based on different appearance of species.

In order to solve these complex problems, the deep convolutional neural network-based approaches are proposed. Our method uses the density map as the ground truth to train the modified classic deep neural networks for object counting regression. Experiments are conducted comparing our proposed models with the state-of-the-art object counting and density estimation approaches. The results demonstrate that our proposed counting model outperforms state-of-the-art approaches by achieving the best counting performance with a mean absolute error of 1.93 and a mean square error of 2.68 on our horticulture nursery plant dataset.

DEDICATION

This study is wholeheartedly dedicated to my family, friends and mentor for their support and help along the way.

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Xiaoning Qian, for his guidance and advice during my Master of Science program. He shows me how to be an independent researcher, and I have learned a lot from his attitude toward work and research. It's a great honor to have him as my advisor.

I would like to thank my committee members, Professor Byung-Jun Yoon, Professor I-Hong Hou, Professor Zhangyang Wang of Texas A&M University, for their helpful suggestions on my research.

I would like to thank Professor Mengmeng Gu at the Department of Horticultural Sciences of Texas A&M University, and Brad Abrameit, Elie Abikhalil, Aaron Cowan of TreeTown USA Wholesale Tree and Plant Farm, for their kind help and support on my research project.

I would liked to thank the professors with their awesome teaching and instructions on the courses I have taken during my study in Texas A&M University, especially Professor Raffaella Righetti, P. R. Kumar, Ulisses Braga-Neto, Xiaoning Qian, Jianer Chen, Shuiwang Ji. I've learned a lot from their courses.

Thanks to my lab mates in Texas A&M University, especially Xiaoqian Jia, Kai He, Qing Jin, Chungchi Tsai, Weizhi Li, Meltem Apaydin, Randy Ardywibowo for their help and encouragement during my study in Texas A&M University.

Finally, I would like to thank my parents and my husband, who have been my source of inspiration and gave me strength, who continually provide their moral, spiritual, emotional and financial support during my Master of Science program.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor Xiaoning Qian, Professor Byung-Jun Yoon, Professor I-Hong Hou of the Department of Electrical and Computer Engineering Department and Professor Zhangyang Wang of the Department of Computer Science and Engineering.

The image data was provided by Professor Mengmeng Gu from the Department of Horticultural Sciences of Texas AM University and Brad Abrameit, Elie Abikhalil, Aaron Cowan of TreeTown USA Wholesale Tree and Plant Farm, Houston, Texas, USA.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by Graduate Merit Scholarship from Texas A&M University.

NOMENCLATURE

CNN	Convolutional Neural Network
MAE	Mean Absolute Error
MSE	Mean Square Error
GLCM	Gray Level Co-occurrence matrices
HOG	Histogram Orient Gradient
UAV	Unmanned Aerial Vehicle
SKU	Stock Keeping Unit
AIHM	AI-Horticulture Monitoring system
ReLU	Rectified Linear Unit
GAN	Generative Adversarial Network

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Machine learning technology has benefited modern human society with the exponential growth of various applications in computer vision, natural language processing, Web search engines, spam detection, and many others [3]. Nowadays, people are using Machine Learning to increase industry productivity, improve business decisions, forecast weather, diagnose diseases, and create many other possibilities.

Among those application topics, estimating the density maps and counting the number of interesting objects from images [4] has gained significant attention in recent years since this topic has a wide range of applications such as crowd counting [5, 6, 2], traffic monitoring [7, 8], cell microscopy in biomedical imaging [9, 10], plant counting in agronomy [11, 12, 13], and environmental surveying [14, 15].

In this project, we focus on the task of counting plants for large-scale breeding nursery farm to build an AI-horticulture monitoring and prediction system using unmanned aerial vehicle (UAV) images. Plant density plays significant roles in decision making influencing crop productivity and sustainability related to many aspects of farming such as watering, fertilizer requirements, and yielding. The traditional manual surveying and decision making process is tedious, time-consuming and prone to human errors especially for large farms. Thus, it is necessary and meaningful to develop alternative (semi-)automatic methods with high-efficiency and accuracy for estimating plant densities.

With the recent development of machine learning methods in computer vision and image analysis, including the traditional ones [16, 17] based on handcrafted image features and more recent deep feature representations derived from training end-to-end deep network architectures [5, 18, 19], we focus on automatic plant counting for this AI-enabled monitoring system in this thesis.

Specifically, the common object counting image analysis challenges are scenario difference, object occlusion, scale variation of views, non-uniform distribution, object appearance difference and perspective difference. In the horticulture plant counting case, another challenge is large vari-

ance of plant species which makes the object features various.

To overcome these challenges, we propose deep neural network models by modifying recent Convolutional Neural Network (CNN) architectures that combine the recent advances in deep learning with classic machine learning methods to automatically count plants.

Extensive experiments are conducted to evaluate our models' effectiveness comparing to recent state-of-the-art object counting density estimation approaches.

In Chapter 2, we will provide a detailed literature review of the previous research on how to count objects in images using Computer Vision and Machine Learning techniques. Chapter 3 provides the UAV-based horticulture monitoring and prediction system setup, experiment design, data collection and preparation. Chapter 4 presents the mathematical model of object counting in images with regression to density maps, and the implementation to address the problem of plant counting in agronomy, including the model analysis and experiment design to evaluate the model effectiveness and accuracy. Potential future research directions are discussed in Chapter 5.

2. LITERATURE REVIEW

One of major components in our proposed AI-Horticulture monitoring system is reliable inventory management. With improved convenience of image collection using UAVs, it is critical to develop customized computer vision and machine learning methods to reliably count the number of plants in specific farm locations. We first review the recent advances in object counting and density estimation in this chapter.

Counting objects and density estimation of objects in images using machine learning techniques have a wide range of applications such as crowd counting [5, 6, 2], traffic monitoring [7, 8], cell microscopy in biomedical imaging [9, 10], plant counting in agronomy [11, 12, 13], and environmental survey [14, 15], which led to an increasing focus by researchers across various fields especially in recent years. This computer vision domain topic comes with many challenges such as object occlusions, distortion of image view, scale differences, non-uniform of illumination and distribution in the image, changing scenarios differences, perspective differences, making the problem difficult to solve. Over the past few years, we have witnessed the considerable development of this topic from earlier approaches of hand-crafted machine learning methods, which are limited with variations of image scale or scene and occlusions of objects, to current state-of-the-art approaches that are more robust and accurate with respect to scale and scene differences.

The researchers have attempted to tackle object counting and density estimation using different methods such as counting by detection, counting by segmentation, counting by regression, and counting by clustering [20]. For regression, the earlier work is using hand-crafted features and the state-of-the-art methods is using Convolutional Neural Networks (CNNs), which have achieved superior performance than the former methods.

In this chapter, we are going to discuss the development of object counting and density estimation from traditional counting approaches to CNN-based approaches. We will also discuss about the methods dealing with the lack of labelled data. Last but not least, the applications on counting plants in agronomy are being discussed.

2.1 Traditional Counting Approaches

Traditional counting approaches often use low-level hand-crafted image features as the input predictor vectors, which are being mapped to object density or count number via different regression techniques. In many surveys of this topic, researchers have categorized the existing methods into detection-based, regression-based, and density estimation-based approaches [21, 20], as detailed below.

2.1.1 Detection-based Approaches

For the detection-based counting approaches, the main idea is using a sliding window going through the whole image to detect objects and count the number of them in an image [22, 23], after hand-crafted feature extraction using Haar wavelets [24], edgelet [25], histogram of oriented gradients [26], or other shape features [27], a classifier is trained via different methods, such as random forests [28], to detect and count objects in images successfully.

However, these detection-based approaches only perform well in images with low object density and clear background. When the objects are crowded together with occlusion, the objects are too small to be detected, or background is cluttered, these approaches have shown limited performance.

2.1.2 Regression-based Approaches

To overcome these issues of detection-based approaches, some researchers have explored the way to do object counting by regression, where a mapping can be learned directly from extracted features to the count number of objects under study [16, 17, 29]. In Idrees [30], the authors have shown that the regression-based approach is more effective than detection-based approaches especially when the image has extremely crowded objects. Generally, counting by regression avoids time-consuming processes related to sliding window detectors. It typically takes two steps: low-level feature extraction and regression modeling. For feature extraction, researchers explored different ways to extract key features [17] such as edge, color, texture (Gray Level co-occurrence matrices-GLCM), gradient (histogram orient gradients-HOG), and so on. For regression model-

ing, multiple regression techniques have been applied and explored, such as linear regression, ridge regression [16], piecewise linear regression [31], to learn the mapping from low-level features to the object counting number in an image. It has been noticed that the traditional regression-based approaches are missing one key feature – the location information of each object we need to count in the image. The traditional regression process is doing a global mapping which ignores the local spatial information.

2.1.3 Density Estimation-based Approaches

In 2010, Lempitsky and Zisserman [4] introduced a new approach to count objects in images by learning a linear mapping from local features to the corresponding density maps indicating both the number of objects in images and the location of each object. By using density maps, it can have both global information about the object counts and local information about object locations without tedious processes needed in detection-based approaches. The integral over the area in the density map gives the object counting results over this area. The learning process is formulated as a convex optimization problem by minimizing a regularized risk quadratic function. This is a milestone as the following research on counting are mostly based on this density map idea.

Knowing that not all the mapping relationships between features to the density map are linear, Pham *et al.* [7] introduced a non-linear mapping learning method via random forest regression to vote for densities of objects. Similarly, Wang and Zou [32] proposed a faster approach based on subspace learning to solve the computational complexity problem in previous methods. However, sometimes the limited feature representations in the existing work may prevent these models from the better counting performance. On the other hand, if adding too many features, the computational complexity makes models not that efficient. In order to solve this trade-off, Xu and Qiu [33] proposed to take richer feature representations into consideration using the random forest as the regression model with the modified tree structures, so that the performance can be boosted with reasonable complexity.

2.2 CNN-based Supervised Learning Approaches

Unfortunately, most of these traditional counting approaches that we have discussed above rely on the expression ability of the hand-crafted extraction features. Recent considerable progress in deep learning and successes of CNNs in computer vision realm have inspired researchers to exploit CNN-based methods to learn the non-liner mapping between image features and the corresponding density map or object counts [5, 34, 7, 2, 18]. Searmanet *et al.* [35] indicated that the features extracted via deep learning models are more effective than the hand-crafted features. The authors in [36] and [5] were among the first ones to apply CNN-base deep learning models to object counting. Wang *et al.* [36] used a modified AlexNet [37] architecture, where the last layer was replaced by a single neuron to predict the object counting number. Zhang *et al.* [5] proposed a cross-scene counting framework to solve the scenario difference problem by training based on two objectives: crowd density and crowd count. Besides from patch-based training, Shang *et al.* [19] proposed an end-to-end training by taking whole images as input and directly outputting the counting results without patch cropping and overlapping.

In order to solve the scale variation problem, there are many scale-aware models such as CCNN and HydraCNN [18] using a pyramid structure, MCNN [2] using a multi-column CNN network. These scale-aware methods can address issues caused by scale differences to some extent, but they are still relying on the models to select scale scopes. More flexible models are needed to tackle these nuisance variation challenges including object appearance and scale variation as well as scenario difference.

2.3 Methods Dealing with the Lack of Labelled Data

The approaches we discussed above mostly are supervised learning with fully annotated ground truth data. In practice, however, labeling the ground truth key points in each image is always labor intensive and expensive. Researchers have made efforts to tackle this issue in counting methods. In 2013, Chen *et al.* [38] proposed a semi-supervised regression framework with ability to perform transfer learning on partially labeled datasets.

Over the past two years, this problem has attracted more attention by researchers. In 2018, Liu *et al.* [39] proposed a novel crowd counting method that leverages abundantly available unlabeled crowd imagery in a learning-to-rank framework. In 2019, Lu *et al.* [40] created a counting model able to count any class of objects by formulating the counting process as a matching problem based on a Generic Matching Network (GMN) architecture to count any objects in any class using the image self-similarity property. Sam *et al.* [41] presented a weakly supervised learning framework by developing Grid Winner-Take-All (GWTA) autoencoder to learn several layers of useful filters from unlabeled crowd images. Wang *et al.* [42] developed a data collector and labeler to automatically generate images and annotate them without any manpower. Olmschenk *et al.* [43] explored generalizing a semi-supervised Generative Adversarial Network (GAN) for object counting.

2.4 Object Counting in Agronomy

Plant counting is a challenging task for today's agronomy. With the increasing demand of plant and crop supplies, it is necessary to perform nursery activities more efficiently and precisely. Usage of remote sensing images can help us to conduct automatic, high-throughput phenotyping and monitoring, for example, to track and manage plant numbers and nursery stages efficiently.

Counting plants from unmanned aerial vehicle (UAV) images using machine learning and computer vision techniques is a new cross-disciplinary application which attracts attention of researchers recently [44, 12, 45]. Machine learning algorithms for detecting and counting an agriculture product with harvesting robots have been applied to grapes, apple, mango, tomato, etc. However, these existing algorithms are designed for high resolution images with either specific species or small plant appearance variations.

The first article related to using UAV image to count plant and crop within individual plots (known as Stock Keeping Units—SKUs) with machine learning techniques [46] was presented in 2018. They have adopted a classic two-step machine learning method to first extract color features in images and then applied machine learning to train a pixel-based segmentation method with two public datasets. This method is a supervised machine learning approach based on the decision tree. Nevertheless, they only used the color features; so this method is limited if the color of plants is similar

with the background or the plants are overlapping. Later on, Oh *et al.* [13] explored approaches using CNN-based deep learning to counting by segmenting sorghum heads with a modified Counting CNN model. This method is also supervised learning and requires significant annotation work. In real-world practice, obtaining the unlabeled data is easy by taking new pictures, but the labeled data are always costly to produce.

3. AI-HORTICULTURE MONITORING SYSTEM IN PRACTICE

In this chapter we briefly introduce our proposed AI-Horticulture Monitoring (AIHM) system and its potential application in horticulture nursery farms. We introduce the real-world setup of such a system in a large horticulture nursery farm in agronomy and discuss the required data preparation for this practical application of counting plants using unmanned aerial vehicle (UAV) images with machine learning and computer vision techniques.

3.1 Background

The thesis research is motivated by the real-world inventory management demand of a local horticulture nursery farm – TreeTownUSA – in Houston, Texas (latitude: $29.33°$, longitude: $-96.20°$). Phenotyping and monitoring plants in the nursery farm such as counting and recognizing plant species can be time-consuming and error-prone, especially for such a large scale nursery farm with multiple plant species. The alternative approach using the low altitude unmanned aerial vehicle (UAV) with high resolution camera to take pictures of the plants and applying machine learning technologies to do counting automatically is more efficient. In our proposed AI-Horticulture Monitoring (AIHM) system, we pay attention to each individual plot (known as Stock Keeping Unit, SKU, in nursary inventory management) by automatic counting the plant number and recognize the plant species in each SKU. We monitor the changes in each specific SKU for inventory management, so that people can make appropriate decisions based on these changes on plant species and number by routine drone flight data collection. In order to build the AIHM system, the work flow is first to have the UAV flight under a set plan with reasonable overlapping, then stitching the images into the panorama whole-view image, and segment individual SKUs, and finally apply the trained machine learning models for counting/species classification of each SKU.

3.2 Field Experiments and Image Acquisition

The drone (UAV) images have been taken at TreeTownUSA during the time of the growing seasons from 2017 to 2019. More than a hundred varieties of plant species are growing in the

nursery farm. Here in our project, we mainly focus on three fields inside the nursery farm: the "West field" with the plant species of mainly oaks with the size around 95 acres; the "Area1" with multiple plant species and the size of round 200 acres, and "Area2" with around 12 acres of multiple plant species.

We can see the figures below that for the plant in West field, it is much easier to count than for the plants in Area1 and Area2. Since the main species is oaks, the appearance between each objects in images does not vary much so that they may have similar feature patterns. However, there are still challenges such as the occlusion of objects, shadows, and view distortion from the UAV images, and when the growing season is different, the species might vary a lot. The main challenges include: (a) huge appearance differences between plant species. The appearance varies in size, color, texture, shape, pose, etc.; (b) self-occlusion of plant objects, especially for some herbaceous plants; (c) super dense distribution for some plant species, which make the manual annotation process difficult. These challenges and problems are what we are going to address in our proposed models.

Figure 3.1: Image taken by UAV on the West field of oak trees.

Figure 3.2: Image taken by UAV on the Area2 with multiple plant species.

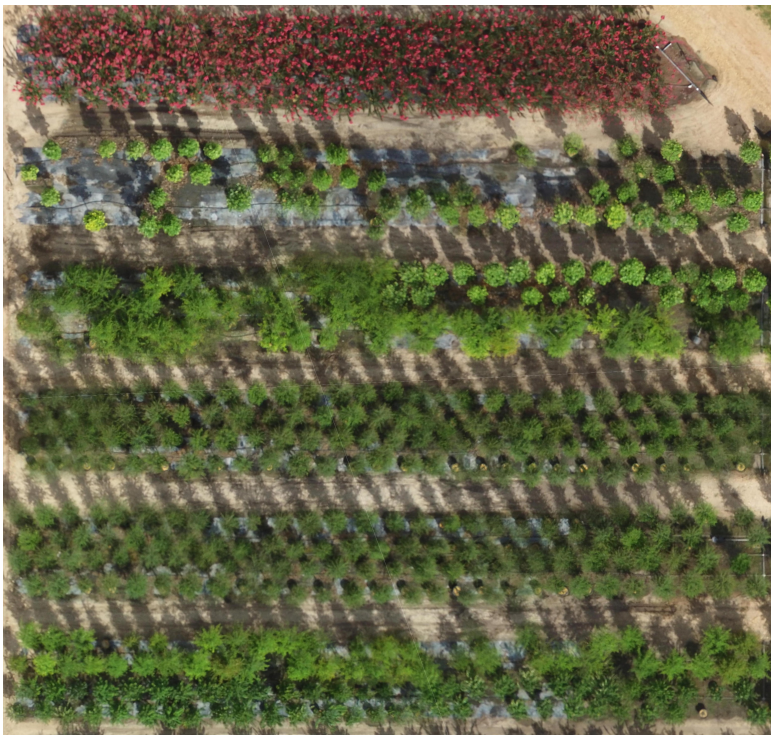

Figure 3.3: Image taken by UAV on the West field with multiple plant species in growing season. (rotated and cropped)

In this thesis, images were captured by high resolution cameras using a low-altitude unmanned aerial vehicle (UAV) at difference heights (60feet, 150feet, 300feet, 400feet) and under different light conditions (cloudy, sunny), with a reasonable overlapping on the flight map to make sure that we could stitch the whole-view map of the fields.

Figure 3.4 shows the various plant species growing in SKUs/plots, our goal is to count the number of plants and at the same time distinguish different plant species in each SKU with accurate, automatic and robust approaches.

3.3 Data Prepossessing of UAV images

3.3.1 Image Stitching

The figures below are the outcomes of whole-view image stitching using a commercial software pix4D. The ortho-mosaic and point clouds were reconstructed for the whole field,

Figure 3.4: The stitched panorama view of the West field. (Area size: 95acres, inside the horticulture nursery farm).

Figure 3.5: The stitched panorama view of the Area1. (Area size: 200acres, inside the horticulture nursery farm).

Figure 3.6: The stitched panorama view of the Area2. (Area size: 12acres, inside the horticulture nursery farm).

3.3.2 SKU/Plot Segmentation

We use the traditional digital image processing approaches to generate the SKU/plots in our target stitched images of horticulture nursery fields. Template matching, image rotation and scaling are applied to find the unique contour of the field. Then the SKU masks are used to extract and segment each SKU area in a fixed order. This can make sure that we generate and number/register each SKU area with the number and order as the same as the previous analysis time, so that by detecting the number of plants and species in each SKU, we could achieve our goal to monitor the difference of each SKU at different growing time of the plant nursery stages.

Figure 3.7: Masked panorama view of the West field.

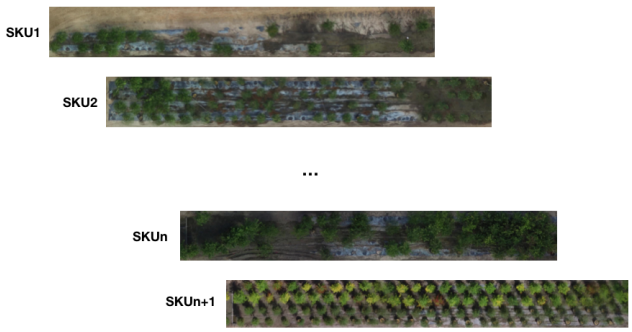

Figure 3.8: Examples of Segmented SKUs/plots.

3.4 Annotations and Density Map Generation

In our thesis, we obtain training labels and evaluation benchmarks by annotating the key points on the objects in images by the Matlab annotation tools, where one object has one corresponding key point, and then the ground truth density maps are generated by applying the Gaussian kernel smoothing on the key points. There are some sophisticated methods about how to generate the ground truth density map [5, 2]. In this study, we use the approach summing a 2D Gaussian kernel centered at each ground truth key point x_{gt} as below:

$$D_i(x) = \sum_{x_{gt} \in S} \mathcal{N}\left(x - x_{gt}, \sigma\right), \tag{3.1}$$

where σ is the scale parameter of the 2D Gaussian kernel and S is the set of all the ground truth key point locations, in our experiments, we choose $\sigma = 15$ as the average size of the plants.

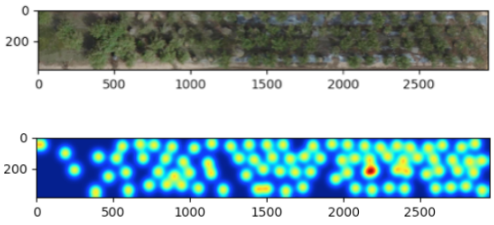

Figure 3.9: The segmented individual SKU with its density map.

4. CNN-BASED OBJECT COUNTING REGRESSION MODELS

In this chapter, several deep neural network models will be implemented to address the problem of plant counting in agronomy. Corresponding experiment design, model prediction results and discussion will be presented to identify the best performing backbone deep network architecture for plant counting.

4.1 Problem Formulation

4.1.1 Learning to Count in Images with Density Maps

The counting objects in image problem can be formulated as a density map estimation problem. We assume that a set of N images $I_1, I_2, ..., I_N$ for training is given, and for each image I_i, each pixel p on it is associated with a feature vector $x_p^i \in \mathbf{R}^K$. For each training image I_i, the objects in it are annotated by a set of 2D key points $\mathbf{P}_i = \left\{P_1, \ldots, P_{C(i)}\right\}$, indicating the location of the objects, where $C(i)$ is the total number of objects in the image. The ground truth density function is defined as a kernel density estimate based on the provided key points[4]:

$$\forall p \in I_i, \quad F_i^0(p) = \sum_{P \in \mathbf{P}_i} \mathcal{N}\left(p; P, \sigma^2 \mathbf{1}_{2\times 2}\right), \tag{4.1}$$

where p denotes any image pixel, $\mathcal{N}\left(p; P, \sigma^2 \mathbf{1}_{2\times 2}\right)$ represents the normalized 2D Gaussian kernel evaluated at image pixel position defined by p. With the density map $F_i^0(p)$, the total object count of number N_i can be obtained by the sum of the ground truth density over the whole image, as follows:

$$N_i = \sum_{p \in I_i} F_i^0(p) \tag{4.2}$$

Note that all the Gaussian are summed, thus the total number of object count is preserved even when there is overlapping between objects in the image, and N_i is close to $C(i)$ since considering the case that some of the points might locate near the boundary.

4.1.2 Basic Mathematical Model for Counting

The first counting model using density maps as the ground truth was illustrated in [4]. Given the set of training images and their corresponding labeled ground truth density maps, the goal is to learn the linear mapping between the feature representation and the density function at each pixel in the image:

$$\forall p \in I_i, \quad F_i(p|w) = w^T x_p^i, \tag{4.3}$$

where x_p^i is the image feature at pixel p and a linear model is assumed with w being the parameter vector to learn from the training data. $F_i(\cdot|w)$ is the estimate of the density function characterized by w. The regularized risk framework is often adopted to find an appropriate w that minimizes the sum of the mismatches between the ground truth and the estimated density functions under regularization:

$$w = \underset{w}{\operatorname{argmin}} \left(w^T w + \lambda \sum_{i=1}^{N} \mathcal{L}\left(F_i^0(\cdot), F_i(\cdot|w)\right) \right), \tag{4.4}$$

where λ is the standard scalar hyperparameter. Once the optimal weight vector has been learned from the training data, the model can predict an estimated density map for an unseen image by a simple linear weighting of the feature vector computed in each pixel in image as in (4.2). So the goal is to solve the above optimization problem by choosing a good loss function $\mathcal{L}$ with the appropriate regularization coefficient λ and computing the optimal w under that loss.

4.1.3 Optimization

The learning model introduced in the last section can be solved by the following convex quadratic program:

$$\min_{w,\xi_1,\ldots\xi_N} w^T w + \lambda \sum_{i=1}^{N} \xi_i, \quad \text{subject to} \tag{4.5}$$

$$\forall i, \forall B \in \mathbf{B}_i : \quad \xi_i \geq \sum_{p \in B} \left(F_i^0(p) - w^T x_p^i\right), \quad \xi_i \geq \sum_{p \in B} \left(w^T x_p^i - F_i^0(p)\right), \tag{4.6}$$

where ξ_i is the auxiliary slack variable and one slack variable ξ_i is for each training image I_i; $\mathbf{B}_i$ denotes the set of subarrays in image I_i. Solving the above quadratic program, the optimal vector

$\hat{w}$ is the solution to (4.3) and the resulting slack variables give the loss $\sum_i \hat{\xi}_i = \mathcal{L}\left(F_i^0(\cdot), F_i(\cdot|\hat{w})\right)$.

Counting objects in images by density maps can be solved with given image features. The remaining challenge is to derive most informative and relevant image features for counting. Furthermore, mappings from the image features to density maps are often not linear. The main objective of our work is to design a more general model that is able to learn not only the linear but also non-linear image features as well as the regression mapping with strong data-driven feature extraction ability from recent deep networks. Motivated by recent successes of data-driven feature engineering by Convolutional neural networks in computer vision, which can extract both linear and non-linear features, in the following sections, we focus on CNN-based regression counting models.

4.2 Basic Model of CNN-based Regression with Density Maps

Given the object counting model with density maps in Section 4.1, the goal of CNN-based model is to learn the non-linear regression function $\mathcal{R}$ that takes an image I as an input, returns the corresponding density map prediction $D_{pred}^{(I)}$,

$$D_{pred}^{(I)} = \mathcal{R}(I|\Omega), \tag{4.7}$$

where Ω is the set of parameters of the convolutional neural network model, for the image $I \in R^{h \times w \times c}$, h, w and c represent the height, width and channels of a given input image.

The typical Convolutional Neural Network (CNN)-based model [18] is shown in the figure below. The training dataset contains the segmented SKUs/plots of plants. For supervised learning, the manual annotation is needed for labeling a key point on each object in the image, then we generate the corresponding density map [2] as discussed previously. The prepared training dataset then is fed into the deep networks to learn a non-linear mapping from the image to an object density map. The last convolutional layer is connected to the following regression loss:

$$l(\Omega) = \frac{1}{2N} \sum_{n=1}^{N} \left\| \mathcal{R}\left(I_n|\Omega\right) - D_{gt}^{(I_n)} \right\|_2^2, \tag{4.8}$$

where Ω represents the neural network parameters, $\mathcal{R}\left(I_n|\Omega\right)$ represents the density map prediction, N is the number of training images, $D_{gt}^{(I_n)}$ denotes the ground truth density for the associated training image I_n.

This is an end-to-end model to directly predict the density map indicating the total number of objects in the image.

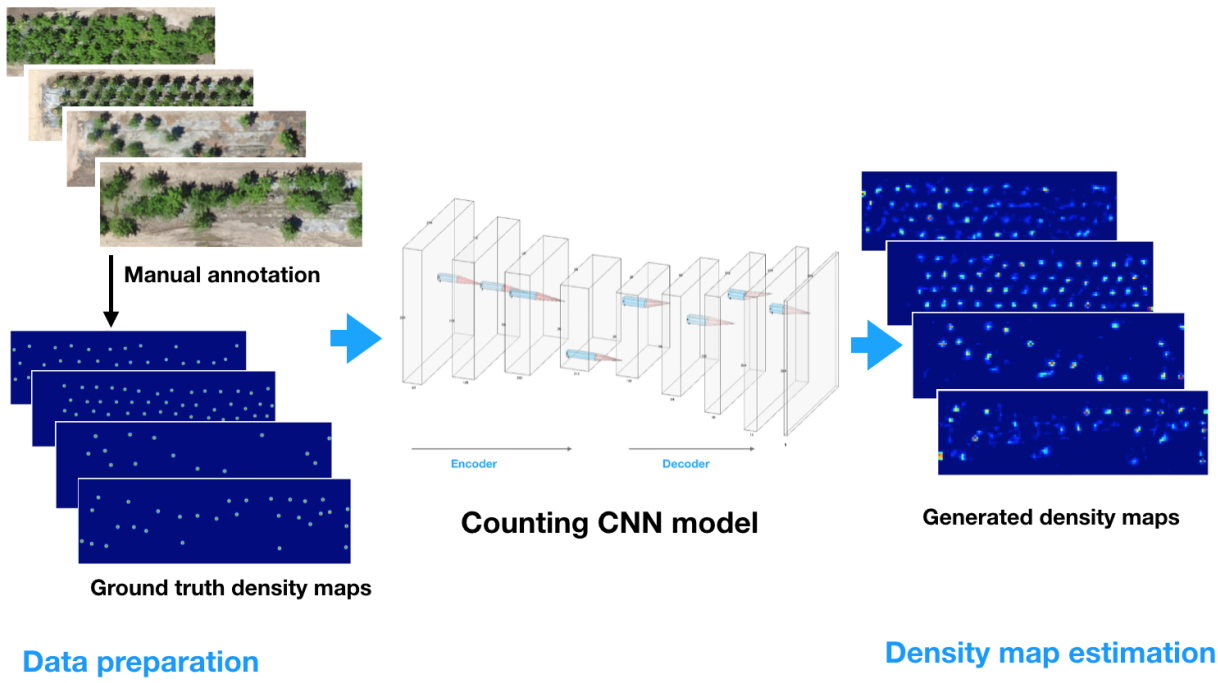

Figure 4.1: Overview of the basic CNN-based regression model with density maps.

4.3 Customized Deep Network Models for Nursery Plant Counting

We now implement, modify, and compare different deep network architectures for plant counting. To estimate the total object number in an image using CNN-based regression model, there are two intuitive ideas: one is a network whose input is the image and the output is the estimated object count number; the other way is to output the density map of the objects, and then obtain the count number by integral of estimated density maps as in (4.2).

In this thesis, we focus on the second choice for the following reasons: 1, The density map preserves more information such as the location of each object, which shows the distribution information; 2, to learn the density map with the CNN models, the learning filters are more adaptive to objects with different size or with different appearance, such as in our horticulture plant counting task with multiple species and different sizes. The second solution strategy is more semantic meaningful and improves the object counting accuracy, as we will see in our experimental results. Our CNN-based regression models will be trained to estimate the density map from an input image. The last layer in the model is compared with the ground truth density map for pixel-to-pixel mapping with the regression loss function (4.8).

For implementing and modifying the CNN-based regression model for counting with density maps, we mainly focus on two things: 1, how to choose the model with strong feature extraction ability; 2, how to improve the accuracy with pixel-to-pixel density map regression. In this section, we discuss about the three classic models: VGG, AlexNet, ResNet can be modified for our horticulture plant counting task.

4.3.1 Modified VGG Plant Counting Model

The overview of the original VGG-16 deep neural network [47] for image classification is illustrated in Figure 4.2, in which there are five sets of convolutional layers with corresponding Rectified Linear Unit (ReLU) and max pooling layers. Two fully connected layers, each with 4096 nodes are then followed by a softmax classifier.

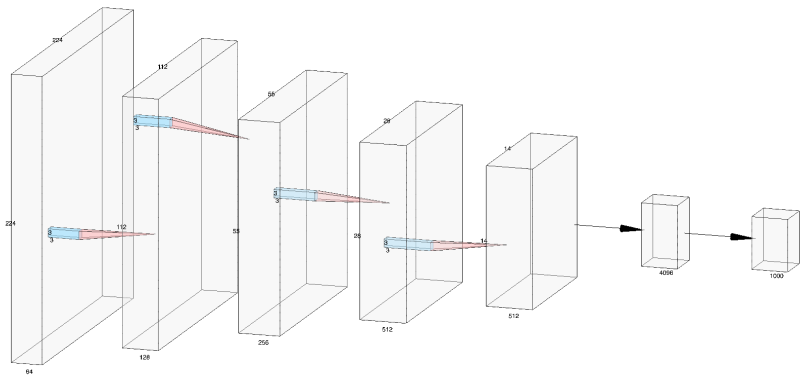

Figure 4.2: Overview of the VGG network architecture.

Based on this VGG-16 deep convolutional neural network model, we have modified it for our nursery plant counting task by taking the first 4 sets of convolutional layers with ReLU and max-pooling layers (10 convolutional layers in total) as the encoder to extract the image features from our horticulture nursery plant images. To enable pixel-to-pixel density map regression, instead of using the remaining layers in VGG-16 including fully convolutional layers, we add three deconvolutional layers and up-sampled by factor of 8 as the decoder. Max pooling is applied for each 2×2 region, and ReLU is adopted as the activation function. This provides the flexible pixel-to-pixel density map regression from extracted image features using the original convolutional layers in VGG-16. The modified VGG-16 model is illustrated in Figure 4.3.

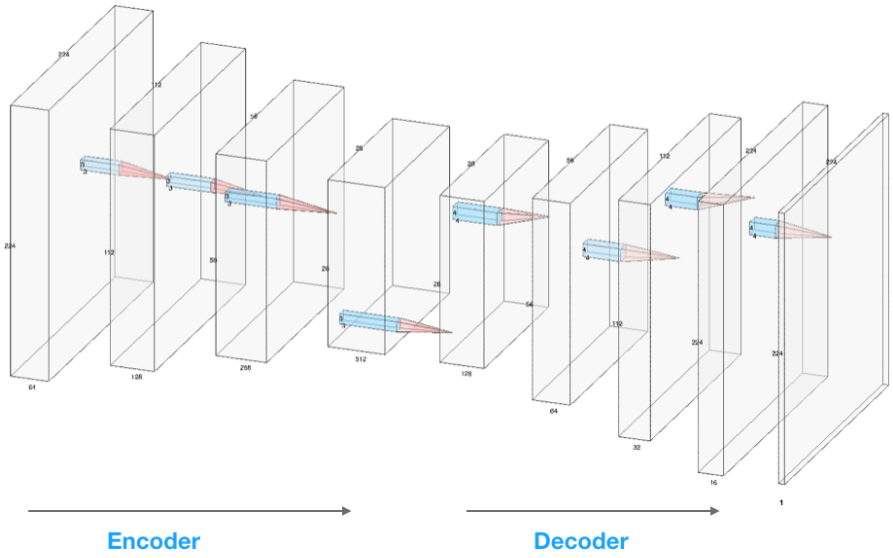

Figure 4.3: The modified VGG model architecture for plant counting.

For the encoder part, after each set of convolutional layers, the image size is scaled to 1/2 of the previous layer. Therefore, after the whole encoder module, the image size is 1/8 of the original input size. In the decoder part, we use three deconvolutional layer with the stride of two pixels, so that the output image size will up-sampled to the original size of the input image, which is also the same size as the ground truth density map. This is indeed critical when we have densely distributed plants in our images. The mean square error (MSE) to measure the difference between the output estimated density map and the ground truth density map is adopted as the loss function as in (4.8) to train this modified density map regression deep network. The optimization for training can be performed by batch-based stochastic gradient descent and backpropagation, as typically done in CNNs.

In recent research of object counting models such as MCNN [2], CSRNet [1], and CCNN [18], as the output image size are down-sampled to 1/4 or 1/8 of the original input image size for the computation consideration, they construct their models to have downsampled density maps to 1/4 or 1/8 of the original image size and compare the downsampled density map with the output predicted density map. However, in this way some of the information in the density map might be

lost. It may affect the accuracy for pixel-to-pixel regression. This issue has inspired us to design the decoder part in the modified VGG model architecture. In our experiments, we have observed that the predicted density map is indeed more accurate for density map regression at the pixel level.

4.3.2 Modified AlexNet Plant Counting Model

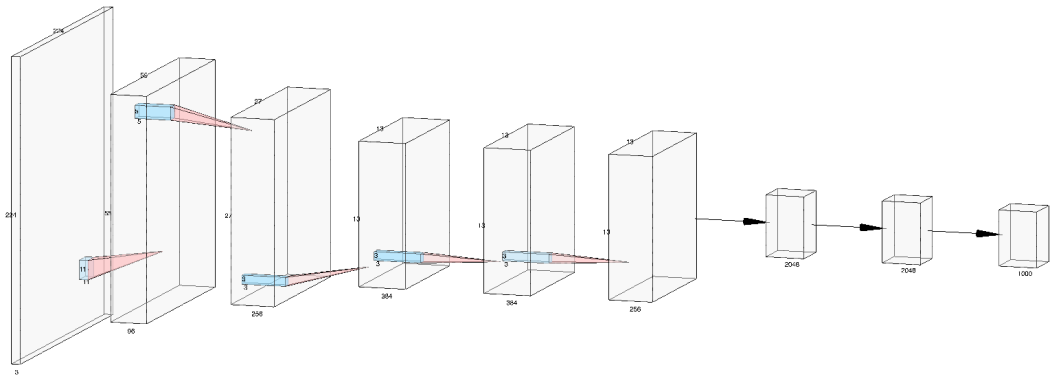

Figure 4.4: Overview of the AlexNet model architecture.

Based on the AlexNet [37] deep convolutional neural network model, we have modified it for our nursery plant counting task by taking off its last three fully connected layers, leaving the first five convolutional layers with pooling as the encoder of our counting model. For the decoder, we use the simple two layers of convolution and up-sampling and returns the one-channel density map. In order to make sure that the feature map size can be divisible, we adjusted the padding size of 4 and 3 on the first two convolutional layers. For the last layer, we use an up-sampling layer with the scale factor of 16 to resize the output of the predicted density map to the original input image size, as discussed in the previous section. The modified AlexNet is presented in Figure 4.5.

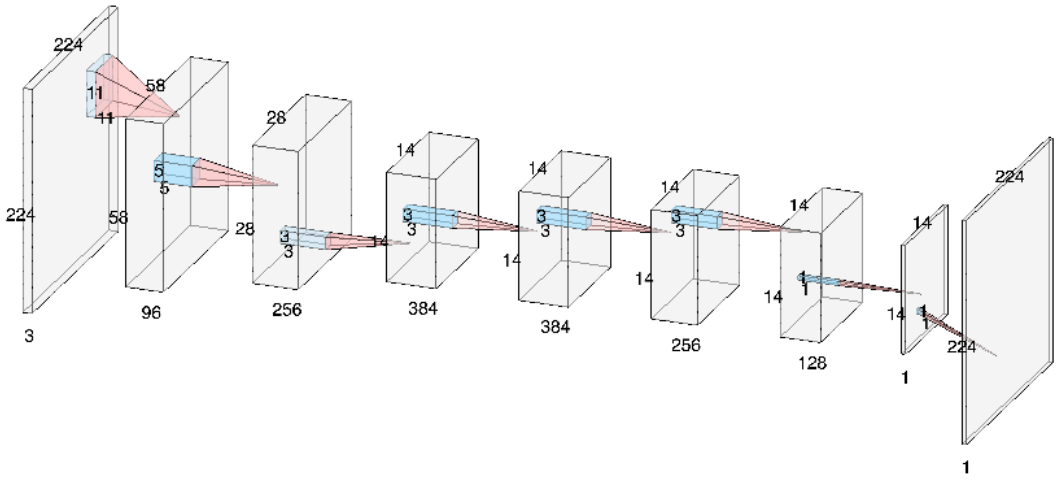

Figure 4.5: The modified AlexNet model architecture.

4.3.3 Modified ResNet Plant Counting Model

Based on the ResNet50 [48] deep convolutional neural network model, we have modified it for our nursery plant counting task by adjusting the stride from 2 to 1 at the layer three to make sure that the output density map is no smaller than the 1/8 of the original image size. We use two convolutional layers as the decoder. In the last layer, we put an upsampling layer of scale factor of 8 to again make sure that output density map is rescaled to the original input image size.

4.4 Evaluation Metrics

The evaluation metrics used to measure the counting performance are (1) mean absolute error (MAE) and (2) mean square error (MSE) based on estimated counts from density maps predicted by different deep models, which are given by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|\,; \tag{4.9}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|^2}, \tag{4.10}$$

where N is the number of test samples; y_i is the ground truth count and $y_i^{'}$ is the estimated count corresponding to the i^{th} sample.

4.5 Implementation Details

All the models are implemented and trained on Texas A&M HPRC (High Performance Research Computing) platform using PyTorch with NVIDIA K80 dual-GPU accelerators. Augmentation techniques are applied such as horizontal flipping, cropping, etc., as typically done in image analysis tasks with deep networks. The modified classic pre-trained models based on ImageNet for VGG, AlexNet, and ResNet are implemented and the parameters in the first convolutional layers in the modified plant counting models are based on these pre-trained models. Our modified models are also compared with several State-of-the-Art CNN-based counting models that are implemented with the default setups as reported in the corresponding papers.

4.5.1 Image Pre-processing

We collect the plant SKU image data via the methods introduced in Chapter 3, mainly from the West field. We manually label them with one dot on each center of the corresponding plant in each SKU image. For object counting, we have generated 670 original SKU images in total. These images are randomly selected and divided into a training set of 400 images and a test set of 270 images. In order to have fair comparison with other CNN-based counting deep networks and make sure that down-sampling layers can have the appropriate output (for example: max-pooling or convolutional layers with stride of 2), we resize our image data size to make it divisible by 16. We resize the SKU image into size 320 x 1024, so that we can do the batch training and save the required memory for training. The ground-truth density maps are scaled at the same rate as the images. Then the density maps are generated based on the average size of nursery plants of $\sigma = 15$ in our UAV images.

4.5.2 Label Normalization

The original generated density maps are generated using the Gaussian kernels with their output values from 0 to 1. The corresponding background pixel value is set to 0. However, we found that the training based on such setup was hard to converge with these networks stuck at local minima. In our reported experimental results, we multiply a factor of 1000 to the original density maps, which helps achieve faster convergence of network training.

4.6 Experimental Results and Discussion

4.6.1 Results and Training Process

Figure 4.6 shows the testing results by one of our modified plant counting models, which is based on ResNet50. We can see that even when the plants are connected to each other with occlusion and shadow in our images, the model can still get a reasonable and accurate counting results.

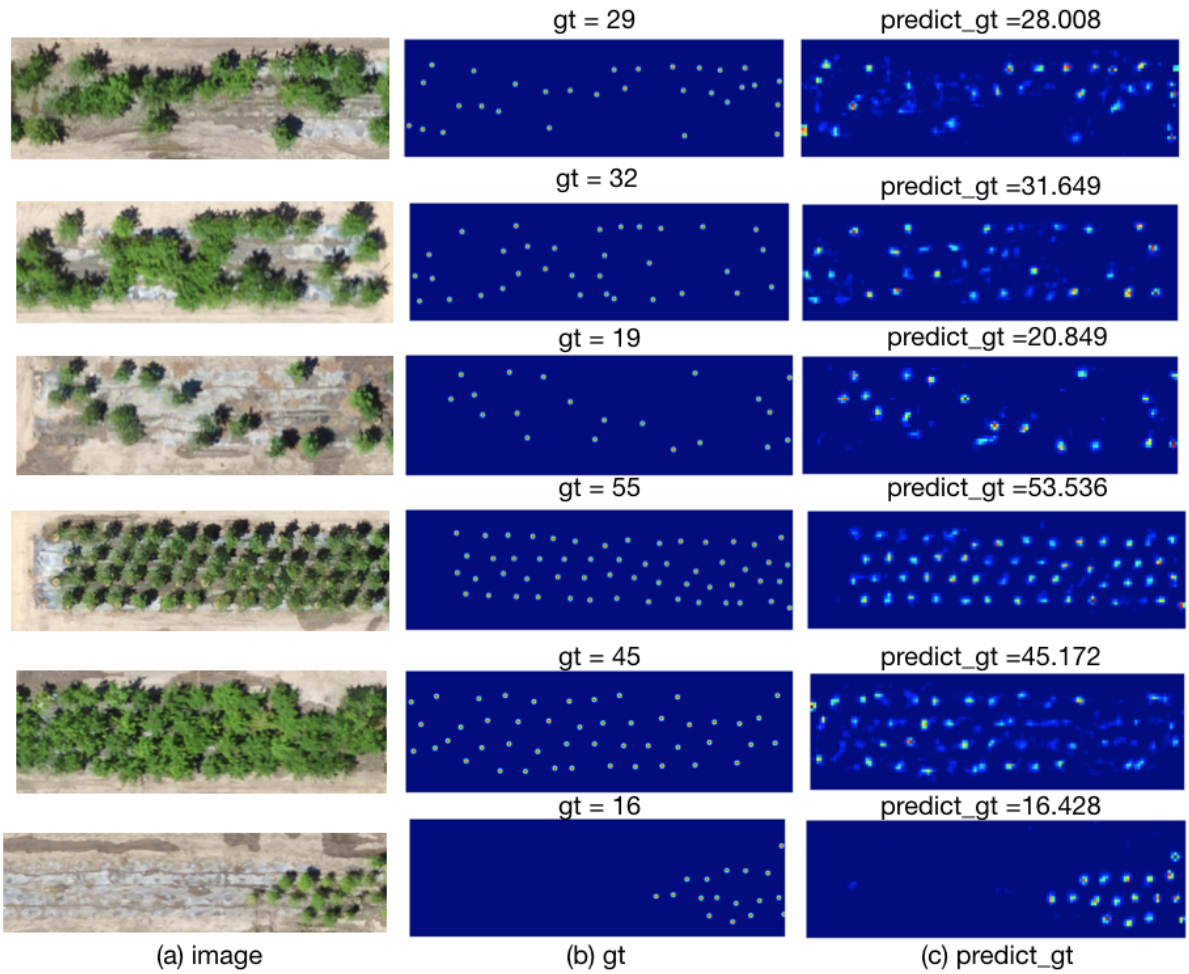

Figure 4.6: Visualization of counting results. Columns: (a) image, (b) ground truth, (c) predicted density map.

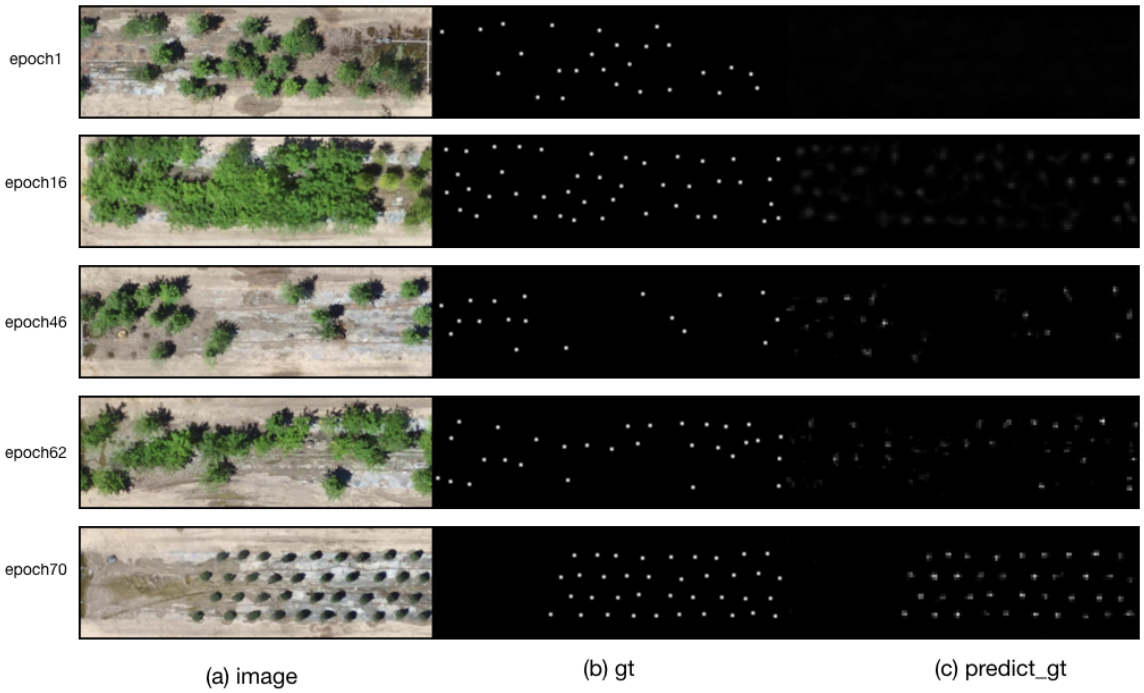

Figure 4.7: The predicted ground truth for one example, of the corresponding training batches during the training process with increasing epochs from the 1st epoch to 70th epoch. Columns: (a) image, (b) ground truth, (c) predicted density map.

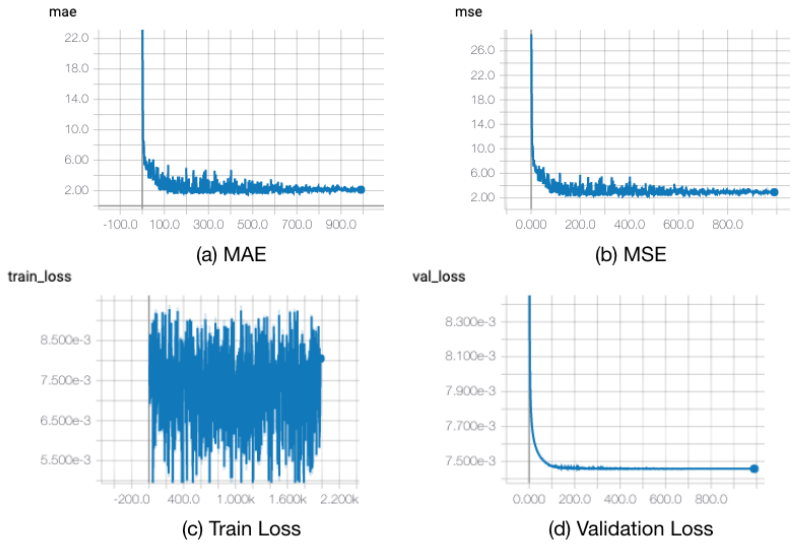

Figure 4.8: The trends of different indices during training for the modified AlexNet model.

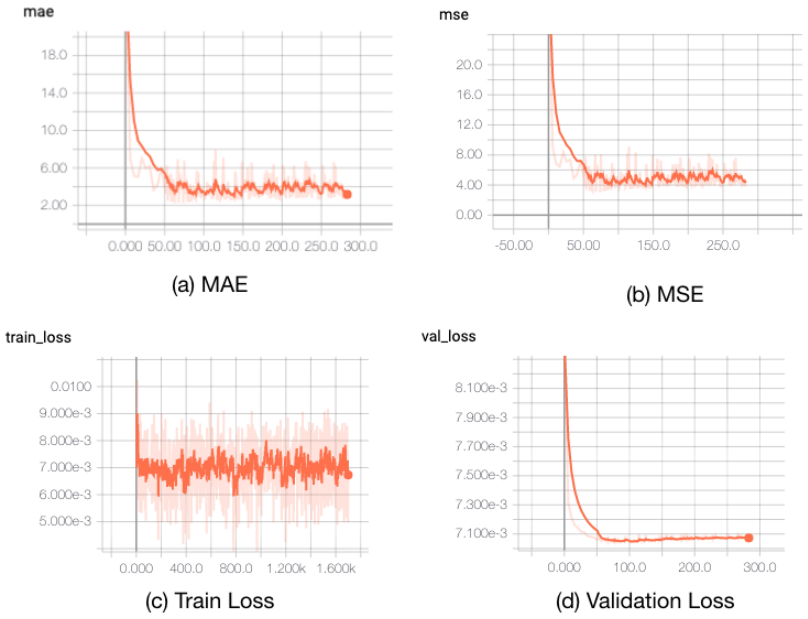

Figure 4.9: The trends of different indices during training for the modified VGG model.

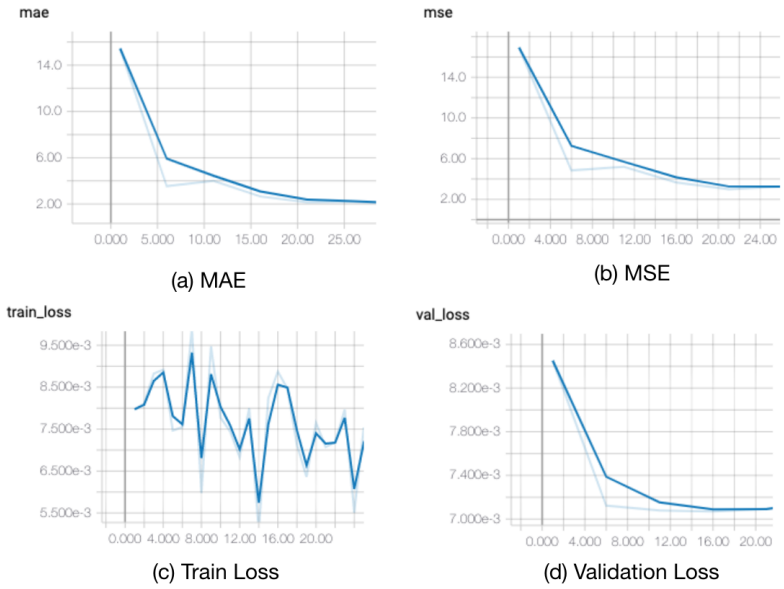

Figure 4.10: The trends of different indices during training for the modified ResNet model.

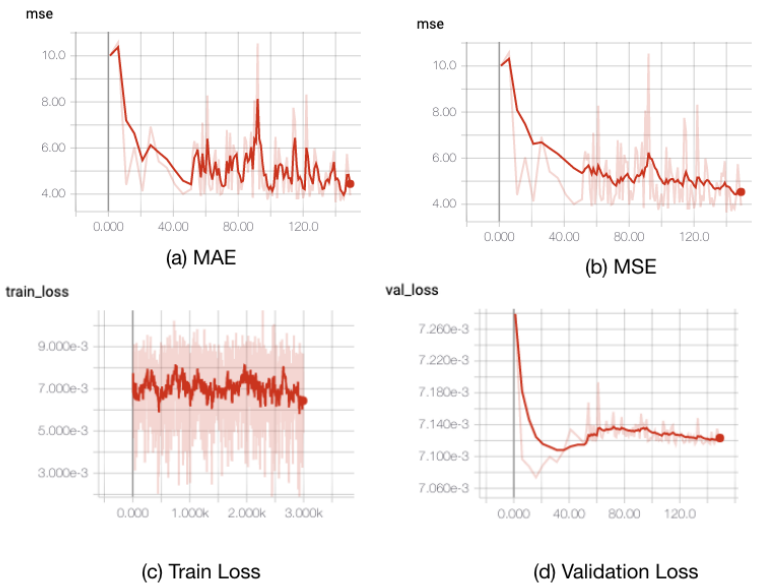

Figure 4.11: The trends of different indices during training for CSRNet [1].

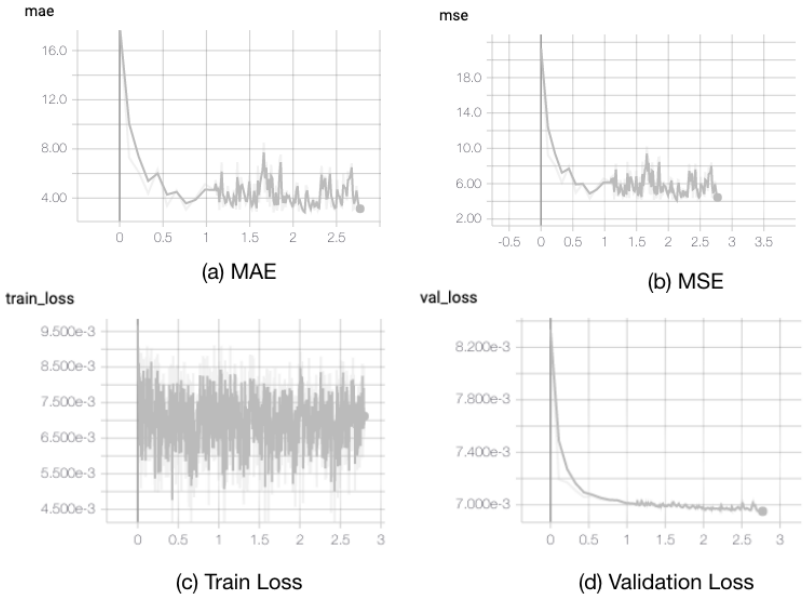

Figure 4.12: The trends of different indices during training for MCNN [2].

4.6.2 Comparing with State-of-the-Arts

Table 4.1 shows the MAE and MSE results of our counting models and the state-of-the-art CNN-based object counting models, including MCNN [2], CSRNet [1], and SANet [49]. Overall, our three modified models with ImageNet pretrained parameters can achieve the state-of-the-art object counting results. The modified AlexNet attains the best results of MAE=1.927 and MSE=2.684, the modified VGG16 model attains the result of MAE = 2.821, MSE = 3.950, the modified ResNet model attains the result of MAE = 2.013, MSE = 2.895.

Table 4.1: Comparison of CNN-based counting models, for plant counting experiments on our UAV images.

Method	MAE	MSE
MCNN	3.470	4.122
CSRNet	2.899	3.941
SANet	3.382	4.540
modified-AlexNet	1.927	2.684
modified-VGG	2.821	3.950
modified-ResNet	2.013	2.895

From the results, we can see that for our dataset, the multi-column design in MCNN, dilated Convolutional Neural Network in CSRNet, and the more complicated multitask SANet are not the most suitable approach for our plant counting dataset. SANet gets over-fitting while training due to the increased model complexity. CSRNet uses VGG16 as the front-end, we can see from Table 4.1 that their results are indeed close to what our modified models achieve. The results demonstrated that the first several convolutional layers of AlexNet, VGG16 and ResNet indeed can extract important image features for our UAV plant images, even they are pre-trained using ImageNet images. They have flexible architectures for modifying and concatenating with customized layers to achieve accurate regression with the ground truth density maps, leading to reliable plant counting results.

5. CONCLUSION AND FUTURE WORK

In this thesis, we have reviewed object counting approaches and implemented CNN-based density estimation for the complex object counting task on images taken by unmanned aerial vehicles (UAVs) in large-scale horticulture nursery farm. The modified models AlexNet, VGG, ResNet can achieve the state-of-the-art accuracy of object counting on horticulture nursery plants.

Applying those models to the real-world applications of plant counting in UAV images and experiments comparing with the state-of-the-art CNN-based counting models MCNN, CSRNet, SANet give us more insight on the effectiveness and limitations of the proposed models.

Modern deep learning models require a large volume of labeled data to be able to generalize well, and applying the model to newly unseen data is dependent on similar training and testing data. When the scenario or object appearance have changed a lot from one dataset to another, new ground truth annotation is always needed to maintain the effective performance of the model. However, because manual annotation is always labor intensive and expensive, we are looking for an alternative method to tackle this issue. Generative adversarial networks (GANs) have shown great potential in semi-supervised learning where the classifier can obtain good performance with very few labeled data [50]. Inspired by this, in the future work, we will explore a semi-supervised method with limited labeled data using the architecture of GAN to explore the possibility in solving this issue.

REFERENCES

[1] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1091–1100, 2018.

[2] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 589–597, 2016.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[4] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in neural information processing systems*, pp. 1324–1332, 2010.

[5] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 833–841, 2015.

[6] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4031–4039, IEEE, 2017.

[7] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3253–3261, 2015.

[8] W. Balid, H. Tafish, and H. H. Refai, "Intelligent vehicle counting and classification sensor for real-time traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 6, pp. 1784–1794, 2017.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," pp. 234–241, 2015.

[10] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 283–292, 2018.

[11] A. K. Nellithimaru and G. A. Kantor, "ROLS: Robust object-level SLAM for grape counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[12] X. Jin, S. Liu, F. Baret, M. Hemerlé, and A. Comar, "Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery," *Remote Sensing of Environment*, vol. 198, pp. 105–114, 2017.

[13] M. h. Oh, P. Olsen, and K. N. Ramamurthy, "Counting and segmenting sorghum heads," *arXiv preprint arXiv:1905.13291*, 2019.

[14] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, "Crowd analysis: a survey," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 345–357, 2008.

[15] G. French, M. Fisher, M. Mackiewicz, and C. Needle, "Convolutional neural networks for counting fish in fisheries surveillance video," *Proceedings of the machine vision of animals and their behaviour (MVAB)*, pp. 7–1, 2015.

[16] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting.," in *BMVC*, vol. 1, p. 3, 2012.

[17] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Crowd counting using multiple local features," in *2009 Digital Image Computing: Techniques and Applications*, pp. 81–88, IEEE, 2009.

[18] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *European Conference on Computer Vision*, pp. 615–629, Springer, 2016.

[19] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 1215–1219, IEEE, 2016.

[20] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd counting and profiling: Methodology and evaluation," in *Modeling, simulation and visual analysis of crowds*, pp. 347–382, Springer, 2013.

[21] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2018.

[22] I. S. Topkaya, H. Erdogan, and F. Porikli, "Counting people by clustering person detector outputs," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 313–318, IEEE, 2014.

[23] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *2008 19th International Conference on Pattern Recognition*, pp. 1–4, IEEE, 2008.

[24] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[25] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 90–97, IEEE, 2005.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005.

[27] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–8, IEEE, 2007.

[28] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011.

[29] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *2009 IEEE 12th international conference on computer vision*, pp. 545–551, IEEE, 2009.

[30] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2547–2554, 2013.

[31] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–7, IEEE, 2008.

[32] Y. Wang and Y. Zou, "Fast visual object counting via example-based density estimation," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3653–3657, IEEE, 2016.

[33] B. Xu and G. Qiu, "Crowd density estimation based on rich features and random projection forest," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–8, IEEE, 2016.

[34] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Fully convolutional crowd counting on highly congested scenes," *arXiv preprint arXiv:1612.00220*, 2016.

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[36] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1299–1302, ACM, 2015.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[38] C. Change Loy, S. Gong, and T. Xiang, "From semi-supervised to transfer counting of crowds," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2256–2263, 2013.

[39] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7661–7669, 2018.

[40] E. Lu, W. Xie, and A. Zisserman, "Class-agnostic counting," in *Asian Conference on Computer Vision*, pp. 669–684, Springer, 2018.

[41] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, "Almost unsupervised learning for dense crowd counting," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, HI, USA*, vol. 27, 2019.

[42] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8198–8207, 2019.

[43] G. Olmschenk, Z. Zhu, and H. Tang, "Generalizing semi-supervised generative adversarial networks to regression using feature contrasting," *Computer Vision and Image Understanding*, 2019.

[44] F. Gnädinger and U. Schmidhalter, "Digital counts of maize plants by unmanned aerial vehicles (UAVs)," *Remote sensing*, vol. 9, no. 6, p. 544, 2017.

[45] S. Liu, F. Baret, B. Andrieu, P. Burger, and M. Hemmerle, "Estimation of wheat plant density at early stages using high resolution imagery," *Frontiers in Plant Science*, vol. 8, p. 739, 2017.

[46] W. Guo, B. Zheng, A. B. Potgieter, J. Diot, K. Watanabe, K. Noshita, D. Jordan, X. Wang, J. Watson, S. Ninomiya, *et al.*, "Aerial imagery analysis quantifying appearance and number of sorghum heads for applications in breeding and agronomy," *Frontiers in plant science*, vol. 9, p. 1544, 2018.

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

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[49] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 734–750, 2018.

[50] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in neural information processing systems*, pp. 2234–2242, 2016.