Attention-Based Deep Bayesian Counting For AI-Augmented Agriculture

Yucheng Wang
wangyucheng@tamu.com
Texas A&M University
College Station, Texas, USA

Mengmeng Gu
mengmeng.gu@colostate.edu
Colorado State University
Fort Collins, Colorado, USA

Mingyuan Zhou
mingyuan.zhou@mccombs.utexas.edu
University of Texas, Austin
Austin, Texas, USA

Xiaoning Qian
xqian@ece.tamu.edu
Texas A&M University
College Station, Texas, USA

Abstract

Object counting in images has been studied extensively, in particular using deep network models recently. The existing counting models typically output the point estimates of the object counts in given images. However, none of these can provide reliable uncertainty quantification of the derived count estimates, which is critical for consequent decision making when adopting these counting models in real-world applications. In this paper, we propose a novel deep counting model in a Bayesian framework. With the designed Bayesian attention module and Bayesian counting loss function, our deep Bayesian counting model not only improves the accuracy of count estimates with varying object and background appearance; but also enables their uncertainty quantification. We specifically focus on plant counting, which plays important roles in AI-augmented agriculture, for example crop yield estimates and farm management. Our ablation studies and experiments with the real-world agriculture data in the Global Wheat dataset have demonstrated that our deep Bayesian counting model obtains high count estimation accuracy as well as reliable uncertainty quantification. In addition, with the integrated Bayesian attention modules, it may help improve the interpretability of the derived count estimates, especially when the distribution of the interested plants in images is heterogeneous.

CCS Concepts

• **Computing methodologies** → **Computer vision tasks**; • **Applied computing** → **Agriculture**; • **Mathematics of computing** → *Bayesian computation.*

Keywords

Object Counting, Bayesian Attention, Uncertainty Quantification

SenSys '22, November 6–9, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9886-2/22/11…$15.00
https://doi.org/10.1145/3560905.3568417

ACM Reference Format:
Yucheng Wang, Mengmeng Gu, Mingyuan Zhou, and Xiaoning Qian. 2022. Attention-Based Deep Bayesian Counting For AI-Augmented Agriculture. In *The 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22), November 6–9, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3560905.3568417

1 Introduction

Object counting based on density map estimation (DME) has diverse computer vision applications in event detection and daily decision making. For example, reliable object counting can help estimate the number of people in rallies [1, 2, 11, 14, 16, 20–25, 28, 30–32], traffic volume in transportation [10, 27], and cell counts in biomedical microscopy images [19, 34]. It is especially useful when the objects we are interested in the given images are highly dense and overlapped.

One key challenge that recent DME methods face is that both the object and background appearances can vary significantly across different images. The variability could exist in object shape, scale, resolution, and objects can appear with different background. In crowd counting, for example, the images might be taken in different distance and different angle, causing the changes of object shape, scale, and appearance. If the objects in a specific image are either larger or sparser than the others, often these end-to-end DME methods with deep neural networks (DNNs) may not be able to adapt due to their inherent limited receptive fields, leading the higher biased estimated counts in that area.

This paper focuses on the application of automated object counting in agriculture applications, where the deployment of drones or unmanned aerial vehicles (UAVs) to monitor growing fields in farms and ranches is becoming commonplace [9, 12, 17]; however, these UAV-captured site images also pose unique challenges to accurate and reliable counting. For example, if a trained model is adopted for different subspecies or different growth stages, it may not be able to produce an accurate prediction.

One important focus of this paper is to enable uncertainty quantification in counting. In agriculture, counting is required so that better decision making regarding planting, fertilizing, irrigation, and other farm management for example, can be derived to help minimize the cost and risk while maximizing the potential yields. In addition to accurate count estimates, it is desirable to also have reliable uncertainty estimates so that robust and sustainable decision making can be achieved when uncertainty arises [6] due to

possible data noise, abnormal UAV image quality, the limitation of adopted machine learning (AI/ML) models, or when deploying AI/ML models that are trained using different data sources, especially considering the challenges of collecting annotated UAV images with the specific crops or plants of interest to different growing fields, farms, or ranches.

We propose a novel uncertainty-aware deep counting method. To achieve accurate and robust object counting, it is important to tackle challenges in analyzing UAV plant images, especially considering the significant variability in object and background appearance, as well as image quality. What's more, with the designed Bayesian attention module and Bayesian counting loss function, our deep Bayesian counting model not only improves the accuracy of count estimates but also enables their uncertainty quantification, without relying on additional deep ensemble model samples as done in previous works [26, 29]. Our ablation studies and experiments with the real-world agriculture data have demonstrated that our deep Bayesian counting model obtains high count estimation accuracy as well as reliable uncertainty quantification. In addition, with the integrated Bayesian attention modules, it may help improve the interpretability of the derived count estimates, especially when the distribution of the interested plants in images is heterogeneous.

2 Related works

2.1 Object Counting

Density-map-estimation (DME) based object counting, proposed first by [19], predicts a density map of a given image. Each pixel value of the density map can be interpreted as the estimated probability of having the object in the corresponding image region. The number of objects can then be calculated by integrating over the density map. More recently, convoluational neural network (CNN)-based DME methods have been proposed by [5], demonstrating its superior performance over traditional object counting methods based on handcrafted features. The counting performance has been further improved to achieve the state-of-the-art (SOTA) performance. Multi-branch CNN proposed in [37] can capture the scale variance. End-to-end deep DME methods [21, 24] utilized bounding box predictions while generating estimated density map predictions. To deal with the issues due to the lack of labelled data, recent research efforts have also been made to explore unsupervised, weakly-supervised or semi-supervised object counting methods using unlabelled or partially labelled data [22, 23, 30, 33]. Those CNN-based DME methods, though mainly aimed at solving the crowd counting problems, can also be applied for vehicle counting [10, 27], counting in cell microscopy images [19, 34], and remote sensing [7]. Some recent works [26, 29] also attempted to develop crowd counting models with the uncertainty quantification capability. In [26], the authors modelled the decomposed uncertainty of the derived crowd density by bootstrap ensembles. In [29], the authors further developed an active sample selection strategy guided by the quantified uncertainty to reduce the amount of labeled data for training. In this work, we focus on the uncertainty quantification of wheat-head counting. Unlike the previous works based on sampled ensembles [26, 29], we directly learn the Bayesian posterior of a modified Bayesian attention module to quantify the uncertainty of the predicted density map for counting.

2.2 Attention Mechanisms in Deep Learning

Attention mechanisms can put different weights to corresponding features to further refine the derived feature maps and highlight features that are important to help make better model predictions. In many computer vision tasks, attention mechanisms are introduced to refine the extracted image features at different levels, capturing long-term dependence and dealing with the limited receptive field of CNNs. Residual Attention modules [35] insert an encoder-decoder network in the residual branch to generate an attention map. SENet [15] generates channel attention weights by pooling the image features over the spatial dimension and recalibrates the derived features. Convolutional Block Attention Modules (CBAM) [36] add a spatial attention map to SENet and thus can further refine the derived features. Those methods have achieved good performance on image classification, detection and semantic segmentation tasks.

Many recent research efforts have been made to apply attention mechanisms to object counting [11, 14, 16, 21]. In [21], a regression-based density map and a detection-based density map are jointly learned with an attention network. In [14], the authors used both global and local attention branches to scale the whole density map and finetune pixel values in local image regions respectively. In [11], the features extracted by applying convolution operations with different dilation rates are fused to enlarge the receptive field and capture features at different scales. In [16], the input image is segmented into sparse and dense regions and the count estimates are derived in these regions respectively. Although these methods can deal with the image appearance variation and achieve good counting accuracy, most of them are trying to use an attention branch to focus on their assigned regions, specifically, dense or sparse ones, to adaptively estimate the corresponding object counts. Some of those attention networks require to be trained separately.

3 Uncertainty-aware counting

In this section, we introduce our Bayesian counting model for agriculture. We first introduce our formulation of the counting problem in Section 3.1. Then in Section 3.2 we discuss how we modify Bayesian attention module and efficiently parameterize the attention module. Lastly in Section 3.3, we describe how we can quantify the counting uncertainty.

3.1 Problem Formulation

Given an image $\mathbf{I}$ with $\{\mathbf{x}_j \in \mathbb{R}^2,\ j = 1, 2, \ldots, J\}$ denoting the corresponding pixel location in the domain of $\mathbf{I}$, let $\{\{(y_n, \mathbf{z}_n)\}_{n=1}^{N_\mathbf{I}}\}$ be the corresponding labels of the $N_\mathbf{I}$ objects of interest in $\mathbf{I}$, where $\mathbf{z}_n$ denotes a point position and $y_n = n$ is the corresponding label. Our objective is to find a mapping $f(\cdot)$ from the given image $\mathbf{I}$ to a density map $\mathbf{D}_\mathbf{I}$. The estimated number of objects in $\mathbf{I}$ can then be calculated by integrating this density map over the image domain: $N_\mathbf{I}^{est} = \sum_{j=1}^{J} \mathbf{D}_\mathbf{I}(\mathbf{x}_j)$. This mapping is often modeled by a deep neural network, denoted by $f_\theta(\cdot)$ where θ is the network model parameters. To train the neural network by backpropagation, we need a loss function to measure the difference of the prediction $f_\theta(\cdot)$ on our training samples to the ground-truth counting label.

One of the most popular loss function is the mean-square error (MSE) loss. Assume that the ground-truth density map of interesting objects in $\mathbf{I}$ and annotated object locations $\{\mathbf{y}_n\}_{n=1}^{N_\mathbf{I}}$, denoted as $\mathbf{D}_\mathbf{I}^{gt}$, can be modelled by the summation of 2-D Gaussian functions with the mean at $\mathbf{z}_n$ and the variance σ^2 corresponding to each object in $\mathbf{I}$:

$$\mathbf{D}_\mathbf{I}^{gt}(\mathbf{x}_j) = \sum_{n=1}^{N_\mathbf{I}} \mathcal{N}(\mathbf{x}_j; \mathbf{z}_n, \sigma^2), \tag{1}$$

where the Gaussian function $\mathcal{N}(\mathbf{x}_j; \mathbf{z}_n, \sigma^2)$ models the probability of the $n-$th object appearing at the corresponding pixel locations in $\mathbf{I}$. More recently, the authors in [25] have proposed a novel Bayesian loss function for counting, combining local constraints in a Bayesian framework. Given $\mathbf{I}$ and its corresponding labelled object locations $\{\mathbf{y}_n\}_{n=1}^{N_\mathbf{I}}$, the Bayesian loss for each training image $\mathbf{I}$ is calculated as:

$$\mathcal{L}^{Bayes}(\mathbf{I}) = \sum_{n=1}^{N_\mathbf{I}} \Big| 1 - \frac{\mathcal{N}(\mathbf{x}_j; \mathbf{z}_n, \sigma^2)}{\sum_{j=1}^{J} \mathcal{N}(\mathbf{x}_j; \mathbf{z}_n, \sigma^2)} \mathbf{D}_\mathbf{I}^{est}(\mathbf{x}_j) \Big|. \tag{2}$$

As we have discussed in Section 1, the variability in object and background appearance can be a great challenge to derive reliable count prediction. The self-attention module, which has been shown to be capable of highlighting the fine-detailed features, could possibly be an effective way of addressing such a problem in object counting. Unlike other existing works that model attention weights to be deterministic, inspired by the recently developed Bayesian attention module [4], we model attention weights as random variables to enable a Bayesian counting framework.

Assume we are given a dataset $\mathcal{D}$ which contains images $\{I_a\}_{a=1}^{A}$, point labels $\{\{(y_n, \mathbf{z}_n)\}_{n=1}^{N_{I_a}}\}_{a=1}^{A}$, and let M denote the predicted weight of self-attention module. We formulate the counting problem as learning the variational distribution $q_\phi(M)$ parameterized by a neural network. Minimizing the Kullback-Leibler (KL) divergence between this variational distribution and the true posterior $D_{KL}(q_\phi(M)||p(M|\{I_a\}_{a=1}^{A}, \{\mathbf{D}_{\mathbf{I}_a}^{gt}\}_{a=1}^{A}))$ is equivalent to maximizing the evidence lower bound (ELBO) given the data, which has the following expression:

$$ELBO = \mathbb{E}_{q_\phi}(\log p_\theta(\{\mathbf{D}_{\mathbf{I}_a}^{gt}\}_{a=1}^{A}|M, \{I_a\}_{a=1}^{A})) - D_{KL}(q_\phi(M)||p_\eta(M)),$$

where θ is the parameter of the density prediction model and $p_\eta(M)$ is the prior distribution of M parameterized by η. Assume the factorized Gaussian likelihood of the density map given I_a and attention weight M, $p_\theta(\mathbf{D}_{\mathbf{I}_a}^{gt}(\mathbf{x}_j)|I_a, M) = \mathcal{N}(f_\theta(\mathbf{x}_j, \mathbf{I}_a, M), \sigma'^2)$, the likelihood of the density map of the dataset $p_\theta(\{\mathbf{D}_{\mathbf{I}_a}^{gt}\}_{a=1}^{A}|M, \{I_a\}_{a=1}^{A})$ can be written as:

$$\begin{aligned} &p_\theta(\{\mathbf{D}_{\mathbf{I}_a}^{gt}\}_{a=1}^{A}|M, \{I_a\}_{a=1}^{A}) \\ &= \prod_{a=1}^{A}\prod_{j=1}^{J} p_\theta(\mathbf{D}_{I_a}^{gt}(\mathbf{x}_j)|I_a, M) \\ &\propto \prod_{a=1}^{A}\prod_{j=1}^{J} \exp(-(\frac{(\mathbf{D}_{\mathbf{I}_a}^{gt}(\mathbf{x}_j) - f_\theta(\mathbf{x}_j, \mathbf{I}_a, M))^2}{2\sigma'^2})). \end{aligned} \tag{3}$$

We derive the negative ELBO, which is our minimization objective as follows:

$$\begin{aligned} \mathcal{L}^{MSE} = &\sum_{a=1}^{A}\sum_{j'}^{J}(\frac{(\mathbf{D}_{\mathbf{I}_a}^{gt}(\mathbf{x}_j) - f_\theta(\mathbf{x}_j, \mathbf{I}_a, M))^2}{2\sigma'^2}) \\ &+ D_{KL}(q_\phi(M)||p_\eta(M)). \end{aligned} \tag{4}$$

Under the independence assumption in [25] and Laplacian likelihood, we can similarly derive the negative ELBO corresponding to the Bayesian loss as follows:

$$\begin{aligned} \mathcal{L}^{Bayes} = &\sum_{a=1}^{A}\sum_{n=1}^{N_{I_a}}(\frac{|1 - \sum_{j=1}^{J} p(y_n|\mathbf{x}_j) f_\theta(\mathbf{x}_j, \mathbf{I}_a, M)|}{b}) \\ &+ D_{KL}(q_\phi(M)||p_\eta(M)). \end{aligned} \tag{5}$$

3.2 Parameterization of Self-Attention Modules

We adopt Convolution Block Attention Module (CBAM) [36] as our self-attention module. Given the feature map at layer l, F^l, the attention weights T of the intermediate activation F^l at layer l can be written as:

$$T(F^l) = \Phi_2^l(\text{ReLU}(\Phi_1^l(F_{\text{pool}}^l))), \tag{6}$$

Φ_1^l and Φ_2^l are the weights of two fully connected layers, and F_{pool}^l denotes the feature map F^l pooled along the width and height dimension. The attention weights M is further derived by applying a sigmoid activation function on each entry of T.

Similar as [4], we model the intermediate attention weights S as random variables whose distribution is parameterized by ϕ, and the distribution of attention weights M is implicitly defined by S. Keep the definition of T in (6), we model S to be the Weibull random variables: $q_\phi(S) \sim \text{Weibull}(k, \frac{ReLU(T)}{\Gamma(1+1/k)})$ and a Gamma prior $p_\eta(S) \sim \text{Gamma}(\alpha, \beta)$ where k, α and β are hyper-parameters. The KL-divergence between a Weibull distribution and a Gamma distribution has the following closed form:

$$\begin{aligned} &D_{KL}(\text{Weibull}(k, \lambda)||\text{Gamma}(\alpha, \beta)) \\ &= \frac{\gamma\alpha}{k} - \alpha\log\lambda + \log k + \beta\lambda\Gamma(1 + \frac{1}{k}) - \gamma - 1 - \alpha\log\beta + \log\Gamma(\alpha). \end{aligned}$$

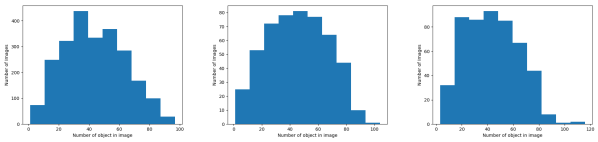
As in [4], instead of directly minimizing $D_{KL}(q_\phi(M)||p_\eta(M))$ in (3) and (5), we alternatively minimize the $D_{KL}(q_\phi(S)||p_\eta(S))$, and the $D_{KL}(q_\phi(M)||p_\eta(M))$ will be implicitly minimized.

3.3 Inference and Uncertainty Quantification of Object Counts

The posterior distribution of the density map given I_a has the following expression:

$$p(\mathbf{D}_{I_a}^{gt}|I_a) = \int p_\theta(\mathbf{D}_{I_a}^{gt}|M, I_a) p_\phi(M|I_a) dM. \tag{7}$$

We infer the density map $\mathbf{D}_{I_a}^{gt}$ by sampling from $p_\theta(\mathbf{D}_{I_a}^{gt}|M, I_a)$ and $p_\phi(M|I_a)$. To quantify the uncertainty of the distribution $p(\mathbf{D}_{I_a}^{gt}|I_a)$, here we consider the variance of the estimated density maps $\mathbf{D}^{est}$. Consider the estimated density map of I_a, we can get multiple samples of the density map of I_a $\mathbf{D}_{I_a}^1, \mathbf{D}_{I_a}^2, \cdots \sim p(\mathbf{D}_{I_a}^{gt}|I_a)$. The variance



(a) Training Set (b) Validation Set (c) Testing Set

Figure 1: Histograms of wheat head counts in Global Wheat Dataset.

of the density map of I_a at pixel x_j is estimated by:

$$\mathrm{Var}_{Map}(I_a;\mathrm{x}_j) = \mathrm{Var}(\mathbf{D}_{I_a}(\mathrm{x}_j)), \mathbf{D}_{I_a} \sim p(\mathbf{D}^{gt}_{I_a}|I_a). \tag{8}$$

To use one number to represent the uncertainty of the test image, we integral over the variance of the density map:

$$\mathrm{Var}^{est}(I_a) = \sum_{\mathrm{x}_j=0}^{J} (\mathrm{Var}_{Map}(I_a;\mathrm{x}_j)). \tag{9}$$

4 Experiments

We evaluate our Bayesian counting model on the Global Wheat dataset [3]. To deal with the challenge due to insufficient annotated training image data, we augment the training set using random cropping and random flipping. We compare our Bayesian counting method with the baseline models. Ablation studies are performed to validate the effect of each model component. To evaluate the counting accuracy and uncertainty estimation reliability, we calculate the mean error and variance of the predicted counts. In the following sections, we first detail the experimental setups and then present our experimental results with discussion.

4.1 Global Wheat Dataset

The Global Wheat Dataset [3] is a large-scale dataset for benchmarking wheat head detection and count estimation. It contains 4,700 high resolution images and 190,000 wheat head labels. In our experiments, we only focus on estimating the number of wheat heads in each image. In 3,373 images that have annotated wheat heads and counts openly accessible to the public, we randomly select 2,362 images for training, 506 images for validation, and 505 images for testing.

Figure 1 illustrates the wheat head count distributions of our training, validation, and testing images. The training images contain on average 43.59 wheat heads, with the standard deviation 20.13. The validation images contain on average 45.20 wheat heads, with the standard deviation 21.20. The test images contain on average 43.59 wheat heads, with the standard deviation 20.58.

4.2 Backbone Architecture and Training Details

We adopt the ResNet18 backbone [13] as the baseline architecture. We remove the last two residual blocks and fully connected layers of ResNet-18, and change the stride of the 5th block of ResNet18 to 1, following [8]. The decoder network contains two 1×1 convolution layers [8] to capture image features. The output is upsampled by 8 to match the size of input images. The network is implemented on PyTorch based on [8]. We use the Adam optimizer [18] and set the learning rate to be 1e-5. The batch size is set to be 25.

4.3 Counting Accuracy Evaluation

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are two widely used evaluation metrics for object counting, which are defined as follows:

$$MAE = \frac{1}{A}\sum_{a=1}^{A}|N_{\mathbf{I}_a} - N^{est}_{\mathbf{I}_a}|, \tag{10}$$

$$RMSE = \sqrt{\frac{1}{A}\sum_{a=1}^{A}|N_{\mathbf{I}_a} - N^{est}_{\mathbf{I}_a}|^2}, \tag{11}$$

where $N_{\mathbf{I}_a}$ is the number of object in $\mathbf{I}_a$, and A is the total number of images. The term $N^{est}_{\mathbf{I}_a}$ in (10) and (11) is the estimated number of objects in $\mathbf{I}_a$, which is calculated by the integral over the whole density map:

$$N^{est}_{\mathbf{I}_a} = \sum_{j=1}^{J}\mathbf{D}^{est}_{\mathbf{I}_a}(\mathbf{x}_j). \tag{12}$$

4.4 Ablation Studies

Our ablation experiment design can be split into three parts. We first evaluate different loss functions and data augmentation methods. Then we add attention modules. Finally, we evaluate the effect of Weibull shape parameter k for stochastic attention modules on three different settings. We will explain the details of each experiment below.

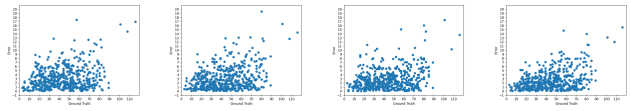
Loss function and data augmentation We experimentally compare the counting performance of Bayesian loss to pixelwise Euclidean loss and evaluate the effect of data augmentation. For Bayesian loss, we set σ to 20. For Pixel-wise Euclidean loss, we set σ to be 10. To help the neural networks trained using pixelwise Euclidean loss to converge correctly, we magnify the ground truth density map by 10.

To speed up the training procedure, we resize all the training images and test images to 512×512. To augment the training data, we randomly select 50% training images and crop them to 512×512, and resize the remaining training images to 512×512. We also flip the training images horizontally and vertically.

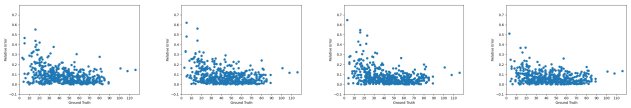
Experimental results can be summarized in Table 1. We further plot the ground truth $N_{\mathbf{I}}$ with respect to the absolute errors ($|N_{\mathbf{I}} - N^{est}_{\mathbf{I}}|$) and relative errors ($\frac{|N_{\mathbf{I}}-N^{est}_{\mathbf{I}}|}{N_{\mathbf{I}}}$) of each test images in Figure 2. We can see that the absolute errors are higher when the test images contain highly dense wheat heads, while the relative errors are higher when wheat heads in test images are sparse. The network trained using Bayesian loss can produce more accurate count prediction than pixelwise Euclidean loss. Data augmentation can improve the counting accuracy on both pixelwise Euclidean loss and Bayesian loss. We observe that with data augmentation, the model will make better count prediction especially on images which the networks trained without data augmentation. This suggests that on highly varying images, the data augmentation is a critical part for training an accurate and robust counting model. In our final model, we train it with Bayesian loss and data augmentation.

Description	RMSE	MAE
ResNet18 + Euclidean loss (w/o aug.)	4.34	3.26
ResNet18 + Bayesian loss (w/o aug.)	3.87	2.88
ResNet18 + Euclidean Loss (with aug.)	4.20	3.13
ResNet18 + Bayesian loss (with aug.)	**3.57**	**2.59**

Table 1: Effect of training loss functions and data augmentation. Counting accuracy is measured by root-mean-square error (RMSE) and mean-absolute error (MAE).



(a) Abs. error with pixelwise loss (w/o aug.) (b) Abs. error with pixelwise loss (with aug.) (c) Abs. error with Bayesian loss (w/o aug.) (d) Abs. error with Bayesian loss (with aug.)



(e) Rel. error with pixelwise loss (w/o aug) (f) Rel. error with pixelwise loss (with aug.) (g) Rel. error with Bayesian loss (w/o aug.) (h) Rel. error with Bayesian loss (with aug.)

Figure 2: Comparison of absolute (abs.) and relative (rel.) errors with respect to the ground-truth counts with different loss functions and augmentation (aug.) setups.

Attention module and data augmentation: Although we have achieved better counting accuracy by integrating with Bayesian loss and data augmentation, the absolute errors on highly dense images and relative errors on sparse images are still high. In this experiment, we evaluate the effectiveness of attention modules for counting results. We compare the counting accuracy of the network models with and without attention modules on two augmentation setups. We insert CBAM attention modules between the 5th and 6th residual blocks, and between the 6th residual block and decoder network. We do experiments to see how attention modules will affect the counting errors of images of different wheat head density.

Similarly, we report our experimental result in Table 2, and plot the ground-truth $N_{\mathbf{I}}$ with respect to the absolute errors ($|N_{\mathbf{I}} - N_{\mathbf{I}}^{est}|$) and relative errors ($\frac{|N_{\mathbf{I}} - N_{\mathbf{I}}^{est}|}{N_{\mathbf{I}}}$) of test images in Figure 4. We observe that the performance are almost the same when the networks are trained without augmented data; however, the counting accuracy improves dramatically by applying attention modules on the augmented dataset. In addition, from Figure 4, we can find that the counting errors reduced on both highly dense images and sparse images. In summary, although the attention modules is a powerful tool that has improved the performance on many other computer vision tasks, we still need to carefully design and train the network to make the best use of them.

Throughout these two ablation studies, we study the effect of loss functions, data augmentation and attention modules on counting accuracy. Our final network architecture design is shown as in Figure 3. We train our Bayesian counting network using Bayesian loss with the aforementioned data augmentation.

Discription	RMSE	MAE
ResNet18 + Bayesian loss (w/o aug.)	3.87	2.88
ResNet18 + Bayesian loss + CBAM (w/o aug.)	3.91	2.85
ResNet18 + Bayesian loss (with aug.)	3.57	2.59
ResNet18 + Bayesian loss + CBAM (with aug.)	**3.19**	**2.33**

Table 2: Ablation studies with attention modules.

k	RMSE	MAE	$3\sqrt{\text{Var}^{est}}$	$2\sqrt{\text{Var}^{est}}$	$1\sqrt{\text{Var}^{est}}$
0.99	3.24	2.36	77.6%	59.1%	33.4%
1	3.23	2.33	92.5%	82.2%	53.6%
5	3.30	2.38	32.0%	22.7%	12.1%

Table 3: Effect of the hyperparameter k of stochastic attention module. Uncertainty estimation is measured by the percentage of the ground-truth in an interval centered at the prediction N^{est} with a bandwidth of six standard deviation $\mathbf{Pr}(N \in N^{est} \pm 3\sqrt{\mathbf{Var}^{est}})$, four standard deviation $\mathbf{Pr}(N \in N^{est} \pm 2\sqrt{\mathbf{Var}^{est}})$, and two standard deviation $\mathbf{Pr}(N \in N^{est} \pm \sqrt{\mathbf{Var}^{est}})$.

Description	RMSE	MAE
ResNet18 + pixelwise Euclidean loss (baseline)	4.34	3.26
ResNet18 + Bayesian loss + Attention	3.19	2.33
ResNet18 + Bayesian loss + Bayesian Attention	3.23	2.33

Table 4: Results on Global Wheat Dataset.

Shape parameter k and different stochastic attention module setups: In the last part of our ablation studies, we study the effect of different stochastic attention modules and Weibull shape parameter k. We model the CBAM attention module between the 6th residual block and decoder network to be stochastic. In the first and second settings, we model the channel attention weights and spatial attention weights as random variables, respectively. In our third setting, we model both the channel attention weights and spatial attention weights as random variables.

We report our experimental results in Table 3. As we can see, modeling the attention module to be stochastic will only slightly degrade the counting performance. The counting accuracy degrades the least when we use stochastic channel attention and set $k = 1$. We also evaluate our uncertainty estimation using the percentage of the cases that the ground-truth counts being within the intervals centered at prediction N^{est} with a bandwidth of $6\sqrt{\text{Var}^{est}}$, $4\sqrt{\text{Var}^{est}}$, and $2\sqrt{\text{Var}^{est}}$. We report the corresponding percentage values in Table 3. With $k = 1$, 92.5% of the predictions are within three standard deviation of the ground-truth counts, indicating reasonable uncertainty quantification performance.

4.5 Results on the Global Wheat Dataset

We have evaluated our uncertainty-aware Bayesian counting model together with the baseline models on the Global Wheat Dataset [3]. The results are reported in Table 4. By incorporating the Bayesian loss, attention modules, and appropriate data augmentation, our model outperforms the baseline models. By introducing stochastic attention weights, we can enable the uncertainty quantification capability of the counting model without significant degradation on counting accuracy.

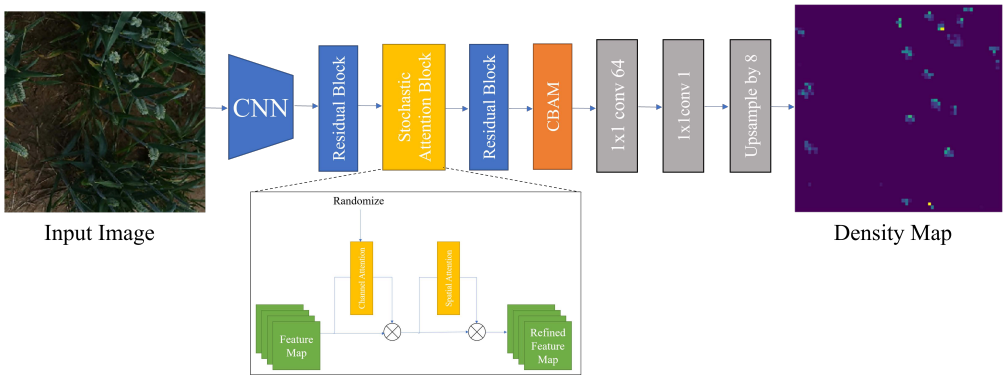
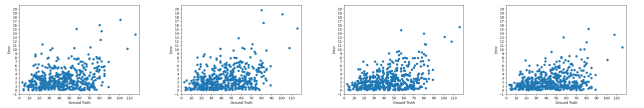
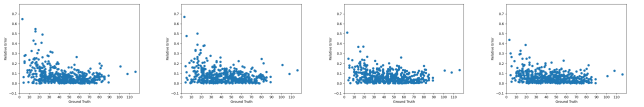


Figure 3: The overall architecture of our proposed model.



(a) Abs. error w/o attention (w/o aug.) **(b) Abs. error with attention (w/o aug.)** **(c) Abs. error w/o attention (with aug.)** **(d) Abs. error with attention (with aug.)**



(e) Rel. error w/o attention (w/o aug.) **(f) Rel. error with attention (w/o aug.)** **(g) Rel. error w/o attention (with aug.)** **(h) Rel. error with attention (with aug.)**

Figure 4: Comparison of absolute (abs.) and relative (rel.) errors with respect to the ground-truth counts with different augmentation setups and with/without (w/o) attention.

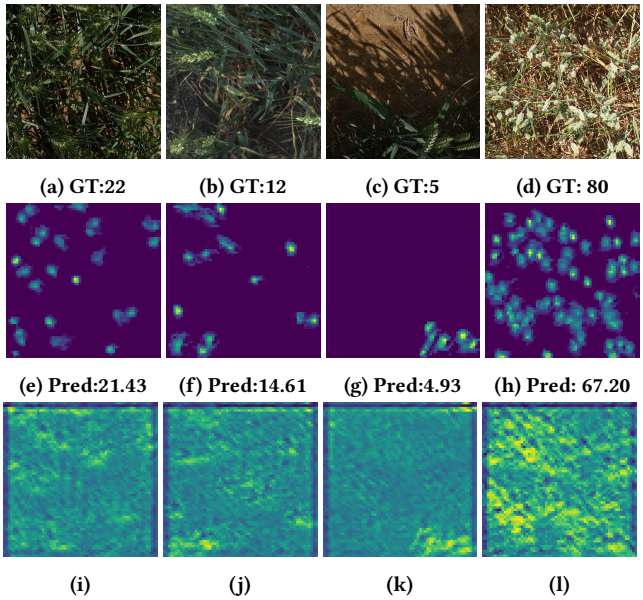
4.6 Visualization and Discussion

In this section, we provide several examples to help visualize the derived predictions by our Bayesian counting model on the Global Wheat Dataset [3]. Figure 5 shows several examples of the test images, density map predictions and attention maps. In the second, third, and forth rows, warmer colors denote higher values while cooler colors denote lower values.

In Figure 5, we can observe that on the Global Wheat Dataset [3], the counting errors are low, even when the background appearance or illumination is complex. In highly dense images, however, the counting accuracy drops significantly. Although the model captures the locations of the most of wheat heads correctly, the model can not give an accurate estimation of the density. Further improvement of the counting accuracy in highly dense plant images will be our future research direction.

5 Conclusions

In this paper, we study object counting in agricultural applications. To improve the performance and tackle the uncertainty issue in object counting, we have introduced attention modules and modelled the attention weights statistically to enable uncertainty quantification in counting, which is the first uncertainty-aware counting method in UAV-captured images to the best of our knowledge. In particular, we use Weibull random variables to model attention weights so that we may derive a distribution of predicted counts instead of only providing point estimates as in many existing object counting models. We evaluate our Bayesian counting model on the Global Wheat Dataset and perform ablation studies to understand the effects of different model setups on counting accuracy with uncertainty quantification. Our experimental results demonstrate that adding attention modules can improve the the accuracy of count estimates, especially when images have varying quality and appearance. More importantly, introducing the randomness to the attention weights enables our first Bayesian counting model with uncertainty quantification, without harming the counting accuracy.



(a) GT:22 **(b) GT:12** **(c) GT:5** **(d) GT: 80**

(e) Pred:21.43 **(f) Pred:14.61** **(g) Pred:4.93** **(h) Pred: 67.20**

(i) **(j)** **(k)** **(l)**

Figure 5: Visualization of test images (5a, 5b, 5c and 5d), count prediction (5e, 5f, 5g and 5h), and Bayesian attention map for uncertainty quantification (5i, 5j, 5k and 5l).

Acknowledgments

This presented work is supported in part by the National Science Foundation (NSF) Awards: CCF-1553281, IIS-1812641, IIS-1812699, CCF-1934904, IIS-2212418, and IIS-2212419.

References

[1] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. 2018. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 734–750.

[2] A. B. Chan, Zhang-Sheng John Liang, and N. Vasconcelos. 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–7. https://doi.org/10.1109/CVPR.2008.4587569

[3] E. David, S. Madec, P. Sadeghi-Tehran, H. Aasen, B. Zheng, S. Liu, N. Kirchgessner, G. Ishikawa, K. Nagasawa, M. A. Badhon, C. Pozniak, B. de Solan, A. Hund, S. C. Chapman, F. Baret, I. Stavness, and W. Guo. 2020. 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. arXiv:2005.02162 [cs.CV]

[4] Xinjie Fan, Shujian Zhang, Bo Chen, and Mingyuan Zhou. 2020. Bayesian Attention Modules. In *NeurIPS 2020: Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2010.10604 (the first two authors contributed equally).

[5] Min Fu, Pei Xu, Xudong Li, Qihe Liu, Mao Ye, and Ce Zhu. 2015. Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence* 43 (2015), 81–88.

[6] Yarin Gal. 2016. *Uncertainty in Deep Learning*. Ph. D. Dissertation. University of Cambridge.

[7] Guangshuai Gao, Qingjie Liu, and Yunhong Wang. 2020. Counting dense objects in remote sensing images. arXiv:2002.05928 [cs.CV]

[8] Junyu Gao, Wei Lin, Bin Zhao, Dong Wang, Chenyu Gao, and Jun Wen. 2019. C^3 Framework: An Open-source PyTorch Code for Crowd Counting. *arXiv preprint arXiv:1907.02724* (2019).

[9] Friederike Gnädinger and Urs Schmidhalter. 2017. Digital counts of maize plants by unmanned aerial vehicles (UAVs). *Remote sensing* 9, 6 (2017), 544.

[10] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. 2015. Extremely overlapping vehicle counting. In *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 423–431.

[11] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. 2019. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1823–1832.

[12] Wei Guo, Bangyou Zheng, Andries B Potgieter, Julien Diot, Kakeru Watanabe, Koji Noshita, David R Jordan, Xuemin Wang, James Watson, Seishi Ninomiya, et al. 2018. Aerial imagery analysis–quantifying appearance and number of sorghum heads for applications in breeding and agronomy. *Frontiers in plant science* 9 (2018), 1544.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/ARXIV.1512.03385

[14] Mohammad Asiful Hossain, Mehrdad Hosseinzadeh, Omit Chanda, and Yang Wang. 2019. Crowd Counting Using Scale-Aware Attention Networks. *CoRR* abs/1903.02025 (2019). arXiv:1903.02025 http://arxiv.org/abs/1903.02025

[15] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2019. Squeeze-and-Excitation Networks. arXiv:1709.01507 [cs.CV]

[16] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. 2020. Attention scaling for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4706–4715.

[17] Xiuliang Jin, Shouyang Liu, Frédéric Baret, Matthieu Hemerlé, and Alexis Comar. 2017. Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery. *Remote Sensing of Environment* 198 (2017), 105–114.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Victor Lempitsky and Andrew Zisserman. 2010. Learning to count objects in images. *Advances in neural information processing systems* 23 (2010), 1324–1332.

[20] M. Li, Z. Zhang, K. Huang, and T. Tan. 2008. 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*. 1–4. https://doi.org/10.1109/ICPR.2008.4761705

[21] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann. 2017. DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. *CoRR* abs/1712.06679 (2017). arXiv:1712.06679 http://arxiv.org/abs/1712.06679

[22] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. 2018. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7661–7669.

[23] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. 2019. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2019), 1862–1878.

[24] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. 2019. Point in, Box out: Beyond Counting Persons in Crowds. arXiv:1904.01333 [cs.CV]

[25] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. 2019. Bayesian Loss for Crowd Count Estimation with Point Supervision. *CoRR* abs/1908.03684 (2019). arXiv:1908.03684 http://arxiv.org/abs/1908.03684

[26] Min-hwan Oh, Peder A. Olsen, and Karthikeyan Natesan Ramamurthy. 2019. Crowd Counting with Decomposed Uncertainty. *CoRR* abs/1903.07427 (2019). arXiv:1903.07427 http://arxiv.org/abs/1903.07427

[27] Daniel Onoro-Rubio and Roberto J López-Sastre. 2016. Towards perspective-free object counting with deep learning. In *European conference on computer vision*. Springer, 615–629.

[28] Z. Qiu, L. Liu, G. Li, Q. Wang, N. Xiao, and L. Lin. 2019. Crowd Counting via Multi-view Scale Aggregation Networks. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. 1498–1503. https://doi.org/10.1109/ICME.2019.00259

[29] Viresh Ranjan, Boyu Wang, Mubarak Shah, and Minh Hoai. 2020. Uncertainty Estimation and Sample Selection for Crowd Counting. *CoRR* abs/2009.14411 (2020). arXiv:2009.14411 https://arxiv.org/abs/2009.14411

[30] Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, and R Venkatesh Babu. 2019. Almost unsupervised learning for dense crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8868–8875.

[31] Sheng-Fuu Lin, Jaw-Yeh Chen, and Hung-Xin Chao. 2001. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 31, 6 (2001), 645–654. https://doi.org/10.1109/3468.983420

[32] Zenglin Shi, Pascal Mettes, and Cees G. M. Snoek. 2019. Counting with Focus for Free. *CoRR* abs/1903.12206 (2019). arXiv:1903.12206 http://arxiv.org/abs/1903.12206

[33] Vishwanath A Sindagi and Vishal M Patel. 2019. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing* 29 (2019), 323–335.

[34] Elad Walach and Lior Wolf. 2016. Learning to count with cnn boosting. In *European conference on computer vision*. Springer, 660–676.

[35] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual Attention Network for Image Classification. arXiv:1704.06904 [cs.CV]

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[37] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 589–597.