The following publication Y. -Y. Ma, Z. -L. Sun, Z. Zeng and K. -M. Lam, "Corn-Plant Counting Using Scare-Aware Feature and Channel Interdependence," in IEEE Geoscience and Remote Sensing Letters, vol. 19, 2022, Art no. 2500905 is available at https://doi.org/10.1109/LGRS.2021.3049489.

Corn-Plant Counting Using Scare-Aware Feature and Channel Interdependence

Yong-Yang Ma, Zhan-Li Sun*, *Member, IEEE*, Zhigang Zeng, *Fellow, IEEE*, and Kin-Man Lam, *Senior Member, IEEE*

***Abstract*—Corn-plant counting is an important process for predicting corn yield and analyzing corn-plant phenotypes. In this paper, an effective corn-plant counting method is proposed, which is based on utilizing the scale-aware contextual feature and channel interdependence. Given the VGG (Visual Graphics Generator) features, the scale-aware features are extracted by spatial pyramid pooling to derive multi-scale context information. In order to utilize the channel interdependent information, the VGG features are integrated via an channel attention module. Moreover, an encoder-decoder structure is constructed to fuse the scale-aware features and the channel interdependence-based features. Considering the sparsity of a corn plant, a hybrid loss function is adopted to train the network, by considering a density map loss function and an absolute count loss function. Experimental results demonstrate the effectiveness of the proposed method for corn-plant counting.**

***Index Terms*—Corn-plant counting, VGG feature, scale-aware feature, channel attention module.**

I. Introduction

Corn-plant counting can provide useful information for agricultural experimentation and the farm management, e.g. breeding programs, predicting corn yield, analyzing corn-plant phenotypes, etc. The most commonly used method is the manual measurement. However, the cost and labor needed are the two most restrictive factors for achieving a high estimation accuracy for a large area [1]. As an alternative, the corn-plant counting can be achieved via the computer vision techniques to reduce the cost and improve the efficiency.

So far, only a few works have been reported for the corn-plant counting, by means of the computer vision techniques. In [1], a convolutional neural networks (CNN) was adopted to segment the images of corn seedlings. Furthermore, the morphological operations and the blob detection are combined to count the corn plants. A digital counting approach for maize plants was proposed in [2] by utilizing the color features of the images captured by an unmanned aerial vehicle (UAV). In [3], a vision-based method was developed to measure corn-plant spacing and population at an early growth stage, by means of vegetation segmentation, image thinning, stem center identification, row line fitting, plant count and plant spacing measurement. A deep convolutional neural network-based approach, termed as TasselNet, was proposed in [4] to accurately counting maize tassels under unconstrained field-based environment.

As part of most related works, many algorithms have been developed for the crowd counting. In [5], a Multi-column Convolutional Neural Network (MCNN) architecture was proposed to accurately estimate the crowd count from a static image with an arbitrary crowd density and arbitrary perspective. A network for Congested Scene Recognition, called CSRNet, was proposed to understand highly congested scenes and perform accurate count estimation [6]. In [7], a scale-aware contextual feature was proposed by adaptively encoding the contextual information of multiple receptive field sizes via an end-to-end trainable deep architecture. An improved scale-adaptive CNN was reported in [8] for the crowd counting. In [9], two modules, namely Spatial-wise Attention Model (SAM) and Channel-wise Attention Model (CAM), were introduced to encode the pixel-wise context of the entire image and to extract more discriminative features from different channels.

Inspired by [7]–[9], in this paper, an effective approach is proposed for the corn-plant counting, by utilizing the scale-aware features and the channel interdependent information. A hybrid loss function, comprising of a density map loss function and an absolute count loss function, is adopted to train an encoder-decoder network structure. Experimental results demonstrate the effectiveness of the proposed method.

The remainder of the paper is organized as follows. In Section II, we present our proposed algorithm. Experimental results and related discussions are given in Section III, and the concluding remarks are presented in Section IV.

II. Methodology

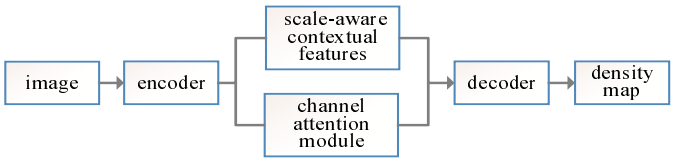


Fig. 1. The flowchart of the proposed counting method for corn plants.

Figure 1 shows the flowchart of the approach proposed for corn-plant counting. There are three main components in the proposed method, including the scale-aware contextual features, a channel attention module, and a hybrid loss function.

The work was supported by a grant from National Natural Science Foundation of China (No. 61972002).

Yong-Yang Ma and Zhan-Li Sun are with School of Electrical Engineering and Automation, Anhui University, Hefei, China.

Zhigang Zeng is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.

Kin-Man Lam is with Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China.

*Corresponding author. (e-mail: zhlsun2006@126.com)

A detailed description of these three parts is presented in the following subsections.

A. Scale-aware contextual features

Figure 2 shows the extraction scheme for the scale-aware contextual features [7]. Given an input image I of a corn plant with a size of $H \times W$, an encoder, which is formed by the first 10 convolutional layers of the VGG-16 Net, is used to extract the local features v_f, with a size of $C \times H \times W$. In terms of spatial pyramid pooling (SPP), v_f are transformed into features with different scales, $k(i) \times k(i)$, by average pooling. For each scale, a 1×1 convolutional layer, consisting of C filters with a size of $C \times 1 \times 1$, is applied to merge the features of C channels. Then, the features are reshaped with the same size as v_f, via upsampling by using bilinear interpolation to obtain the scale-aware features s_i. Given v_f and s_i, a contrast feature c_i is computed as follows:

$$c_i = s_i - v_f. \tag{1}$$

This feature captures the differences between the features at a specific location and those in the neighborhood. Then, a 1×1 convolutional layer, followed by a sigmoid function, is applied on c_i to derive the weight w_i. Considering the saliency, the weighted scale-aware features are computed as,

$$a_f = \frac{\sum_{i=1}^{l} w_i \odot s_i}{\sum_{i=1}^{l} w_i}, \tag{2}$$

where l is the number of scales, and $\odot$ is the element-wise product between a weight map and a feature map. Finally, the scale-aware contextual features c_f are obtained by concatenating a_f and v_f, i.e.,

$$c_f = [a_f \mid v_f], \tag{3}$$

where $[\cdot \mid \cdot]$ denotes the channel-wise concatenation operation.

B. Channel attention module

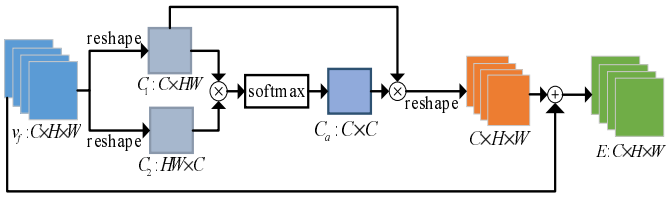


Fig. 3. Architecture of the channel attention module.

Figure 3 shows the architecture of the channel attention module [9]. For v_f, two feature maps C_1 and C_2 are attained by the reshape operation. Then, the channel attention map C_a can be computed by using softmax, as follows:

$$C_a^{nm} = \frac{\exp\left(C_1^m \cdot C_2^n\right)}{\sum_{m=1}^{C} \exp\left(C_1^m \cdot C_2^n\right)}, \tag{4}$$

where C_a^{nm} denotes the m^{th} channel's influence on the n^{th} channel. Finally, the output E^n of CAM is defined as follows:

$$E^n = \lambda \sum_{m=1}^{C} \left(C_a^{nm} \cdot C_1^m\right) + v_f^n, \tag{5}$$

where λ is a parameter learned via a convolutional layer, with a kernel size of 1×1.

C. Hybrid loss function

The density map loss function $L_D(\theta)$ is a commonly used loss function for crowd counting,

$$L_D(\theta) = \frac{1}{M} \sum_{i=1}^{M} \left\| F\left(X_i; \theta\right) - F_i \right\|^2, \tag{6}$$

where θ is the set of parameters to be learned in the network, M is the total number of training images, and X_i is the i^{th} input image. For X_i, F_i and $F\left(X_i; \theta\right)$ are the ground-truth density map and the estimated density map, respectively.

Besides $L_D(\theta)$, the absolute count loss function $L_Y(\theta)$ can also be used to evaluate the estimation performance,

$$L_Y(\theta) = \frac{1}{M} \sum_{i=1}^{M} \left\| Y(X_i; \theta) - Y_i \right\|^2, \tag{7}$$

where $Y\left(X_i; \theta\right)$ and Y_i are the estimated count and the ground-truth count, respectively.

Given $L_D(\theta)$ and $L_Y(\theta)$, a hybrid loss function L can be adopted to jointly optimize the network to improve the generalization ability, defined as follows:

$$L = (1 - \mu) L_D(\theta) + \mu L_Y(\theta), \tag{8}$$

where μ is a weighting coefficient [8].

In Fig. 1, the decoder, consisting of multiple convolutional layers, is adopted to convert the scale-aware features and the channel interdependence-based features into the estimated density map.

III. Experimental Results

A. Experimental data and set-up

The performance of the proposed approach is evaluated on an image set provided by Anhui Agricultural University, Hefei, China. The image set was captured from a part area of an agricultural demonstration base on August 14, 2019. The agricultural demonstration base, covering an area of about 663300 square meters, is located at Bozhou City, Anhui Province, China.

An UAV, Mavic, of SZ DJI Technology Co., Ltd., equipped with a visible light camera, flied about 10-12 meters high above a corn field of the demonstration base. There are 896 images in the data set. For each image, the locations of corn plants are manually marked to obtain the ground-truth data. All of annotated core plant points are convolved by a Gaussian kernel to encode the ground-truth density map.

Under a mild illumination intensity, as illustrated in Fig. 4, we can easily recognize a corn plant. Nevertheless, under a strong illumination intensity, reflection happens on the plants, as shown in Fig. 5. Some areas are easily misclassified as the corn plants. Therefore, two different sets of experiments will be conducted by considering different illumination intensities.

For the corn-plant image sets, the original image size (5472×3648) is relative large. We extract image patches with different sizes from the original images, and use them as the

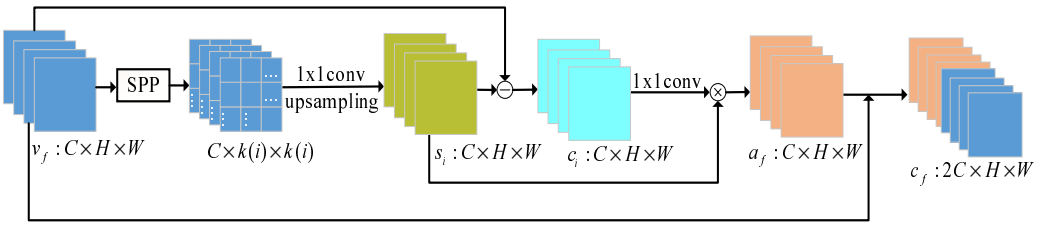


Fig. 2. The extraction scheme for scale-aware contextual features.

Fig. 4. An image of a corn plant captured under a mild illumination intensity.

Fig. 5. An image of a corn plant captured under a strong illumination intensity.

TABLE I
THE RELATED INFORMATION OF THE CORN-PLANT IMAGE SETS IN EXPERIMENTS.

Illumination	Set	N_s	N^A_{max}	N^A_{min}
mild	train	301	158	35
	test	100	144	40
strong	train	364	198	42
	test	131	178	47

training and the testing sets. For these image patches, the smallest size is 730 $\times$ 545, and the largest size is 1312 $\times$ 1054. Table I shows the sample number (N_s) in the training set and the testing set. Moreover, Table I also lists the smallest number (N^A_{min}) and the largest number (N^A_{max}) of the corn plants annotated in these image patches.

Two widely used performance indices, mean absolute error (MAE) and mean squared error (MSE), are adopted to evaluate the estimation accuracy of corn-plant density, defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y(X_i;\theta) - Y_i|, \tag{9}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y(X_i;\theta) - Y_i)^2}, \tag{10}$$

where N is the number of testing images, and $Y(X_i;\theta)$ and Y_i are the estimated count and the ground-truth count for the i^{th} image, respectively.

Moreover, another performance index, mean absolute percentage error (MAPE), is used to evaluate the percentage of the counting errors, defined as follows:

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{Y(X_i;\theta) - Y_i}{Y_i}\right| \tag{11}$$

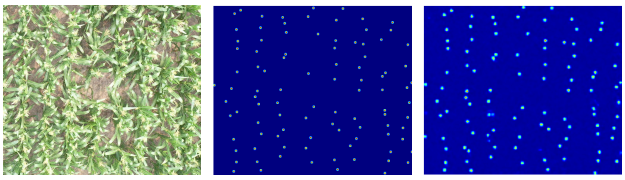
To evaluate the performance of the proposed method, denoted as SC, we compare it with four state-of-the-art algorithms for object counting, including CSRNet [6], SCAR (Spatial-Channel-wise Attention Regression) [9], MCNN [5], and CAN (Context-Aware Network) [7]. All experiments were conducted in the PyTorch environment, running on a Tesla V100 GPU with 32 GB memory.

B. Experimental comparison under mild illumination intensity

TABLE II
THE ESTIMATION ERRORS FOR CORN-PLANT DENSITY UNDER A MILD ILLUMINATION INTENSITY.

Method	MAE	MSE	MAPE
CSRNet	5.33	7.14	6.60%
SCAR	4.59	5.94	5.14%
MCNN	4.29	5.48	5.06%
CAN	2.67	3.66	3.11%
SC	2.38	3.02	2.83%

Under a mild illumination intensity, there are 301 training samples and 100 testing samples. Table II shows the estimation errors for the corn-plant density. We can see that, the estimation errors of the proposed method and CAN are obviously lower than that of other three methods. Compared to CAN, our MAE and MSE are 10.86% and 17.49% lower, respectively. As an example, Figs. 6(a)- 6(c) show a testing image, the corresponding ground-truth corn-plant density, and the estimated corn-plant density, respectively.



(a) Testing image (b) Ground-truth density map (c) Estimated density map

Fig. 6. An example of a testing image, the ground-truth density map, and the estimated density map, under a mild illumination intensity.

C. Experimental comparison under strong illumination intensity

TABLE III
THE ESTIMATION ERRORS FOR CORN-PLANT DENSITY UNDER A STRONG ILLUMINATION INTENSITY.

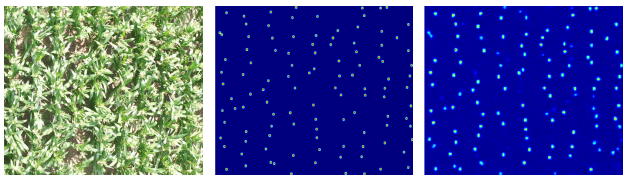
Method	MAE	MSE	MAPE
CSRNet	8.53	10.54	8.75%
SCAR	8.38	10.72	8.35%
MCNN	6.33	8.40	6.53%
CAN	3.69	4.92	3.71%
SC	3.45	4.69	3.41%

There are 364 training samples and 131 testing samples under the strong illumination intensity. Table III shows the estimation errors for corn-plant density. We can see that, compared to the mild illumination intensity, the estimation errors are obviously higher for all the methods under the strong illumination intensity. Thus, a strong illumination intensity has a significant effect on the accuracy of density estimation of corn plants. The estimation errors of the proposed method and CAN are obviously lower than that of the other three methods. Our method has its MAE and MSE at about 6.50% and 4.67%, respectively, lower than CAN. As an example, Figs. 7(a)- 7(c) show a testing image, the corresponding ground-truth density map, and the estimated density map, respectively.

D. Related discussions

In order to evaluate the performance gain achieved by different features and loss functions, we conducted the following experiments, with five combinations of the features and loss functions.

(1) **VGG+DM**: the VGG feature and the density map (DM) loss are used in the network.



(a) Testing image (b) Ground-truth density map (c) Estimated density map

Fig. 7. An example of a testing image, the ground-truth density map, and the estimated density map, under a strong illumination intensity.

TABLE IV
THE CORN-PLANT COUNTING ERRORS WITH DIFFERENT COMBINATIONS OF FEATURES AND LOSS FUNCTION UNDER THE MILD ILLUMINATION INTENSITY.

Method	MAE	MSE	MAPE
VGG+DM	5.33	7.14	6.60%
SA+DM	2.67	3.66	3.11%
SA+CI+DM	2.58	3.43	3.04%
SA+HL	2.52	3.23	3.01%
SC	2.38	3.02	2.83%

(2) **SA+DM**: the scale-aware (SA) feature and the density map loss function are used in the deep neural network, i.e., the CAN method.
(3) **SA+CI+DM**: the scale-aware feature, the channel interdependence-based feature (CI), and the density map loss function are used in the network.
(4) **SA+HL**: the scale-aware feature and the hybrid loss (HL) function are used in the network.
(5) **SC**: the scale-aware feature, the channel interdependence-based feature, and the hybrid loss function are all used in the deep neural network, i.e., the proposed method.

Tables IV and V show the corn-plant counting errors for the above five different combinations, under two different illumination intensities. It can be seen that, compared to the CAN method, i.e., SA+DM, the corn-plant counting errors decreased, when SA is combined with either HL or CI. The performance is the best, when both HL and CI are combined with SA, i.e., the proposed method.

In (8), the term $(1-\mu)L_D(\theta)$ should have a relative larger proportion than $\mu L_Y(\theta)$. Take the mild illumination intensity for example, Table VI shows the corn-plant counting errors when the parameter μ is set as different values. In experiments, the parameter μ is set as 0.1 for the proposed method.

Table VII shows the runtime (sec.) of one forward computation for the various methods. It can be seen that the runtime is close to each other.

IV. CONCLUSION

In this paper, an effective approach is proposed for the corn-plant counting. In the proposed method, the scale-aware features, the channel interdependent information, and the hybrid loss function are combined, and have been verified to be able to effectively strengthen the counting performance for corn plants. Compared to the existing state-of-the-art

TABLE V
THE CORN-PLANT COUNTING ERRORS WITH DIFFERENT COMBINATIONS OF FEATURES AND LOSS FUNCTION UNDER THE STRONG ILLUMINATION INTENSITY.

Method	MAE	MSE	MAPE
VGG+DM	8.53	10.54	8.75%
SA+DM	3.69	4.92	3.71%
SA+CI+DM	3.57	4.86	3.51%
SA+HL	3.55	4.89	3.50%
SC	3.45	4.69	3.41%

TABLE VI
THE CORN-PLANT COUNTING ERRORS WHEN THE PARAMETER μ IS SET AS DIFFERENT VALUES FOR THE PROPOSED METHOD.

μ	MAE	MSE	MAPE
0.001	2.45	3.17	2.92%
0.01	2.49	3.14	2.98%
0.05	2.42	3.15	2.90%
0.1	2.38	3.02	2.83%
0.15	2.40	3.01	2.90%
0.3	2.48	3.19	3.04%

algorithms, the proposed algorithm can achieve a competitive performance.

REFERENCES

[1] B. T. Kitano, C. C. Mendes, A. R. Geus, A. R., H. C. Oliveira, and J. R. Souza, "Corn plant counting using deep learning and UAV images," *IEEE Geoscience and Remote Sensing Letters*, DOI: 10.1109/LGRS.2019.2930549, 2019.

[2] F. Gnadinger and U. Schmidhalter, "Digital counts of maize plants by unmanned aerial vehicles (UAVs)," *Remote Sensing*, vol. 9, no. 6, pp. 544-544, 2017.

[3] C. Wang, X. Guo, and C. Zhao, "Detection of corn plant population and row spacing using computer vision," *Proceedings of the International Conference on Digital Manufacturing and Automation*, 2011, pp. 405-408.

[4] H. Lu, Z. G. Cao, Y. Xiao, B. H. Zhuang, and C. H. Shen, "TasselNet: counting maize tassels in the wild via local counts regression network," *Plant methods*, vol. 13, no. 1, pp. 1-17, 2017.

[5] Y. Y. Zhang, D. S. Zhou, S. Q. Chen, S. H. Gao, and Y. Ma , "Single-image crowd counting via multi-column convolutional neural network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589-597.

[6] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1091-1100.

[7] W. Z. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5099-5108.

[8] J. Sang, W. Wu, H. Luo, H. Xiang, Q. Zhang, H. Hu, and X. Xia, "Improved crowd counting method based on scale-adaptive convolutional neural network," *IEEE Access*, vol. 7, pp. 24411-24419, 2019.

[9] J. Y. Gao, Q. Wang, and Y. Yuan, "SCAR: Spatial-/channel-wise attention regression networks for crowd counting," *Neurocomputing*, vol. 363, pp. 1-8, 2019.

TABLE VII
THE RUNTIME (SEC.) OF THE FORWARD COMPUTATION FOR THE VARIOUS METHODS.

Method	CSRNet	SCAR	MCNN	CAN	SC
runtime	0.19	0.18	0.11	0.18	0.19