SELF-ADVERSARIAL TRAINING AND ATTENTION FOR MULTI-TASK WHEAT PHENOTYPING

Gensheng Hu, Lidong Qian, Dong Liang, Mingzhu Wan

The authors are **Gensheng Hu,** Professor, School of Electronics and Information Engineering, Anhui University, Hefei, Anhui, China; **Lidong Qian,** under postgraduate, School of Electronics and Information Engineering, Anhui University, Hefei, Anhui, China; **Dong Liang,** Professor, School of Electronics and Information Engineering, Anhui University, Hefei, Anhui, China; **Mingzhu Wan**, undergraduate, School of Information Science and Technology, Fudan University, Shanghai, China. **Corresponding author:** Gensheng Hu, No. 111 Jiulong Road, Economic and Technological Development Zone, Hefei, Anhui, China; phone: +86-134-8560-7849; e-mail: hugs2906@sina.com.

ABSTRACT.

Phenotypic monitoring provides important data support for precision agriculture management. This study proposes a deep learning-based method to gain an accurate count of wheat ears and spikelets. The deep learning networks incorporate self-adversarial training and attention mechanism with stacked hourglass networks. Four stacked hourglass networks follow a holistic attention map to construct a generator of self-adversarial networks. The holistic attention maps enable the networks to focus on the overall consistency of the whole wheat. The discriminator of self-adversarial networks displays the same structure as the generator, which causes adversarial loss to the generator. This process improves the generator's learning ability and prediction accuracy for occluded wheat ears. This method yields higher wheat ear count in the Annotated Crop Image Database (ACID) data set than the previous state-of-the-art algorithm.

Keywords.

Plant phenotype, attention mechanism, self-adversarial networks, stacked hourglass.

Research on plant phenotype usually includes plant growth analysis, root phenotypic analysis, disease identification, and leaf and seed enumeration. Image-based plant phenotypic analysis can be classified as holistic or component analyses (Choudhury et al., 2016). Holistic analysis mainly considers the geometric properties of the whole plant. Component analysis considers various parts of a plant, such as spikelets, leaves, stems, and fruits. Wheat is an important crop in the world, and its phenotype analysis is related to wheat yield and food security. This study focuses on the component analysis of wheat ear and spikelet counting, which are important research subjects of wheat phenotype.

In practical application, images of wheat are derived by cameras in wheat fields and wheat ears and spikelets are counted by using image processing algorithms. The result can be used to estimate wheat yield and analyze wheat ear distribution. Shi et al. (2013) developed a clustering algorithm to estimate plant counting, plant location, and interplant spacing. Perrin et al. (2005) presented a framework for extracting tree crowns by minimizing the energy function. The framework produced good results in a relatively short period of time, even in the case of dense plantations. Fernandez–Gallego et al. (2018) proposed a wheat ear counting method for field conditions. Ear counting was determined by using Find Maxima to segment local peaks. Zhou et al. (2018) proposed multi-feature optimization and TWSVM method for wheat ear counting under field conditions. The aforementioned methods adopt classical machine learning or image processing technologies for plant counting. These methods need to manually extract features and expect certain requirements for the shape and color of the targets in the original images.

Deep learning networks, such as DCNN, RCNN, R-FCN, GoogLeNet, ResNet, VGGNet, AlexNet, and GAN, have been used for plant phenotyping in recent years (Hu, 2018; Fuentes, 2017; Singh, 2018; Valerio, 2017). GAN allows unsupervised training of generative models (Goodfellow et al., 2014), but it is unstable and difficult to train. DCGAN, an full convolutional architecture, successfully solved the limitation of GAN (Radford et al., 2015). DCGAN can generate heat maps of labels as in saliency (Luc et al., 2016). Adding the adversarial training strategy to plant phenotyping can bring benefits, such as improved accuracy and recall rate. Visual attention model is effective in computation and image understanding. Existing methods usually use recursive neural networks to generate the attention map of an image area at each step and combine the information of different steps to make a final decision (Bahdanau, 2015; Ba, 2015; Kuen, 2016). Thus far, no study has applied attention model to wheat phenotyping.

This study aims to further explore the state-of-the-art progress in wheat phenotypic research. A deep ConvNet model with attention mechanism is designed to learn the structure of wheat ear phenotypes through adversarial training. The designed networks perform multi-task learning, which simultaneously locate and count ears and spikelets. The contributions of this study are presented as follows:

- A self-adversarial training architecture is applied in wheat phenotypic multi-task learning. A model is developed to solve wheat phenotypic problem by using Deep Convolutional Generative Adversarial Networks (DCGAN) techniques.

- A visual attention mechanism is added to self-adversarial training. This mechanism drives the model to focus on the regions of interest and learn the structure of wheat ears and spikelets.

- The proposed method is evaluated using ACID data set (Pound et al., 2017), and the effects of different components on the design are analyzed. The experimental results show improved F1 score on the data set.

MATERIALS AND METHODS

DATA ACQUISITION AND PRE-PROCESSING

Wheat images for training and validating the proposed algorithm were derived from the annotated crop image database (ACID) (Pound et al., 2017). ACID is the latest date set of wheat plants under glasshouse conditions and contains 534 images, including more than 4,000 annotated ears and 48,000 annotated spikelets. Figure 1 shows the annotation of wheat ears and spikelets. The native resolution of these images is 1956×1530. The occluded spikelets are not annotated, but partially occluded ears are retained. Figure 2 shows a few challenging examples from the ACID data set.

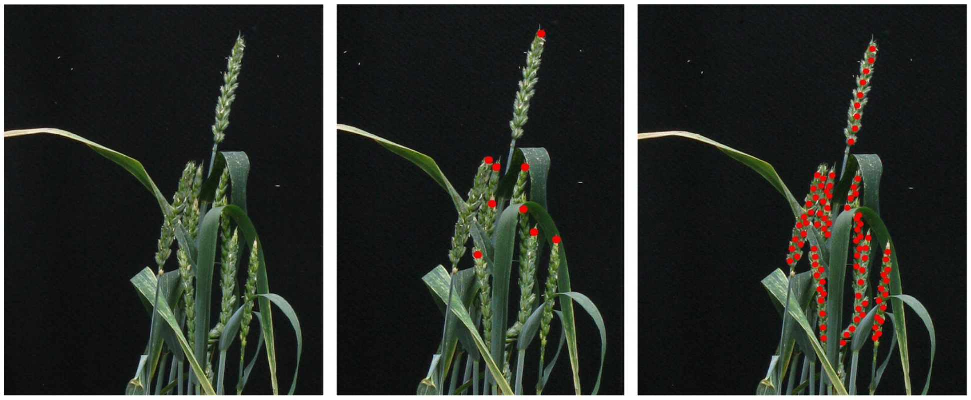

Figure 1. Annotation of wheat ears and spikelets

Deep network architecture restricts the size of the input images due to hardware and software limitations. Each wheat image used in this study is divided into several regions during the training process, and the target ear is centered on the images with roughly the same scale. The default initial cropping is 384 pixels, and the image patches are wrapped to the size of 256×256. The image patches are randomly rotated, scaled and flipped to ensure robustness of the model training.

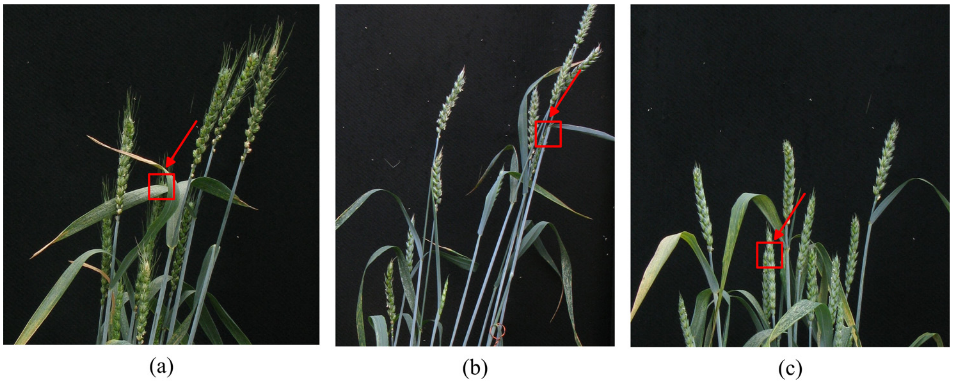

Figure 2. Challenging examples from the ACID data set (a) slight occlusion, (b) severe occlusion, and (c) self-similarity

Self-Adversarial Training with Attention Mechanism

Generative adversarial networks (GAN) consist of a generator and a discriminator (Goodfellow et al., 2014). The generator is used to simulate the original target data distribution and generate target samples. The discriminator is used to determine whether the generated samples are real. These components work adversely to learn the representative features of the given data set. GAN has been used to identify plant diseases and insect pests (Hu, 2019; Zhou, 2019).

The generator used in the proposed self-adversarial networks is a four-stacked hourglass network accompanied by four attention maps. After feeding forward through the first part, the generator will obtain two sets of heatmaps that indicate the confidence scores of wheat ears and spikelets at every location. The discriminator displays the same architecture as the generator. However, the RGB image is encoded separately with the generated heatmaps and the ground-truth heatmaps, which are used as input of discriminator. Figure 3 shows an overview of the proposed network architecture.

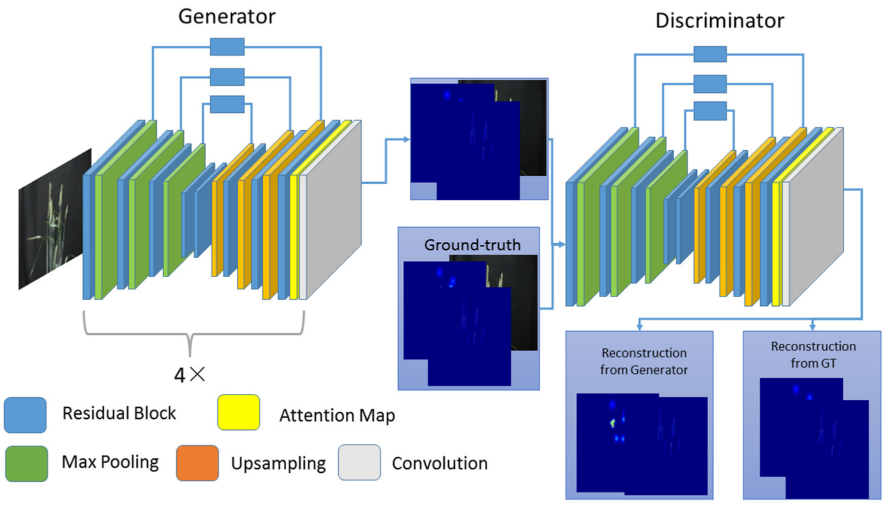

Figure 3. Overview of the proposed network architecture

Generator

The generator uses hourglass network as the base network and provides a repetitive top–down, bottom–up structure. The generator helps capture information from input images at different scales. Many cues, including wheat orientation, spikelet arrangement, and the relationship among adjacent spikelets, can be identified at different scales in the images. Additionally, attention mechanism is introduced and combined with the hourglass network. This process allows the generator not only to learn the features from multiple scales but also focus on the whole wheat. Each blue block in Figure 2 represents one or more residual blocks, which combine convolution, batch normalization, and an additional skip layer to avoid vanishing gradients and help train in this very deep network. All residual blocks output 256 features. Each yellow block in Figure 2 represents an attention map. The generated attention maps are used to reweight the features to automatically infer the regions of interest. The network inputs are represented by 256×256 RGB images. After a series of initial processing, the size of the original image is changed to 256×256, and the output size of the generator's final attention mapping is 64×64×2, where 2 represents two classes of ears and spikelets.

Discriminator

The discriminator distinguishes real data from the generated ones. The discriminator architecture is the same as that of the generator. The inputs of the discriminator correspond to ground-truth or generated heatmaps, which are bound to the corresponding RGB images of the wheat. The discriminator learns whether the locations of ears and spikelets on the heatmaps are correct and correspond to the wheat in the input color images. The discriminator also reconstructs two sets of new heatmaps, which exhibit quality that depends on the similarity between the reconstructed and inputted heatmaps. Discriminator loss corresponds to the error between the inputted and reconstructed heatmaps.

Attention Mechanism

Visual attention is an important mechanism for the brain to effectively understand the scene. Traditional models define regions of interest manually by a set of rectangle bounding boxes. An attention model provides a principled way to focus on target regions by generating attention maps. Attention model has achieved great success in occlusion and self-similar human pose estimation (Chu et al., 2017). Occlusion and self-similarity problems also exist in wheat phenotyping. Therefore, an attention model is introduced into the proposed network architecture. A holistic attention map is used to encode the configurations of the whole wheat within each hourglass network. Conventional attention mechanisms often use global softmax to model spatial correlations. A disadvantage is that the entire image is normalized on the basis of the constant factor, and the local neighborhood spatial correlation is ignored. Conditional random fields (CRF) can be used to model the spatial correlations so attention map-driven networks can concentrate on complex wheat ear configurations (Chu et al., 2017).

In the CRF model, a hard attention method is adopted. The attention label of the i-th position is $y_i = \{0, 1\}$. The probability for $y_i=1$ is obtained iteratively using the mean-field approximation as follows:

$$\Phi(y_i = 1)_t = \begin{cases} \sigma(\psi_u(i)) & t = 1 \\ \sigma(\psi_u(i) + \sum_j w_{i,j}\Phi(y_j = 1)_{t-1}) & t > 1 \end{cases} \tag{1}$$

Where $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function, $\psi_u(i)$ measures the cost of attention label $y_i=1$ which can be derived from feature maps, and $w_{i,j}$ is a filter independent on time steps.

Convolutional features and feature map are denoted as f and s, respectively. Thus,

$$\mathrm{s} = \mathrm{g}(\mathrm{W}^a * \mathrm{f} + \mathrm{b}) \tag{2}$$

Where W^a represents convolution filters, and g is a nonlinear activation function. The attention map Φ_t of stage t can be expressed as follows (Chu et al., 2017):

$$\Phi_t = M(s, \mathrm{W}^k) = \begin{cases} \sigma(\mathrm{W}^k * \mathrm{s}) & t = 0 \\ \sigma(\mathrm{W}^k * \Phi_{t-1}) & t = 1, 2, 3 \end{cases} \tag{3}$$

Where M denotes a weight shared convolution sequence for the mean field approximation, $\sigma(x)$ is the sigmoid function, and W^k is the spatial correlation kernel which is shared at different time steps.

TRAINING

The generator is trained by the back-propagation of its own loss L_{MSE} and the adversarial loss L_{adv} from the discriminator. The loss function L_{G} of the generator is presented as (Chou et al., 2018):

$$L_{\mathrm{G}} = L_{\mathrm{MSE}} + \mu_{\mathrm{G}} L_{\mathrm{adv}} \tag{4}$$

Where μ_{G} is a hyperparameter used to control the weight of the adversarial loss L_{adv}. The generator consists of N stacks of hourglass modules with attention model. The output of each attention model contains M attention maps, each of which contains a Gaussian peak at the ground-truth location of the j-th ear. Every ear contains the locations of the spikelets. The loss L_{MSE} of the generator itself can be expressed as:

$$L_{\text{MSE}} = \sum_{i=1}^{N}\sum_{j=1}^{M}(H_{ij} - \hat{H}_{ij})^2 + \lambda_{\text{m}}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{Q}(H_{ijk} - \hat{H}_{ijk})^2 \tag{5}$$

Where H_{ij} is the ground-truth heatmap of j -th wheat ear on i -th stack, $\hat{H}_{ij}$ is the generated ear heatmap. H_{ijk} is the k -th spikelet of j -th wheat ear on i -th stack in the ground-truth heatmap, $\hat{H}_{ijk}$ is the generated spikelet heatmap, Q is the number of spikelets in each ear, and λ_{m} is a balance parameter used to control the weight loss of spikelets. Using L_{MSE} as a loss function, the generator is enforced to learn image features, which are important for localizing ears and spikelets. Adversarial loss L_{adv} enforces the generator to reduce unreasonable prediction and can be expressed as:

$$L_{\text{adv}} = \sum_{j=1}^{M}(\hat{H}_j - D(\hat{H}_j, X))^2 \tag{6}$$

Where $\hat{H}_j$ is the output heatmap of the generator, X is the input image, and $D(\hat{H}_j, X)$ is the output heatmap of the discriminator. Adversarial loss L_{adv} calculates the error between the generated and reconstructed heatmaps.

The generated and ground-truth heatmaps of ears and spikelets are separately fed into the discriminator for training. Two sets of heatmaps are reconstructed to calculating L_{real} and L_{fake}. Discriminator loss is expressed as:

$$\begin{aligned}
&L_{\text{real}} = \sum_{j=1}^{M}(H_j - D(H_j, X))^2 \\
&L_{\text{fake}} = \sum_{j=1}^{M}(\hat{H}_j - D(\hat{H}_j, X))^2 \\
&L_{\text{D}} = L_{\text{real}} - k_t L_{\text{fake}} \\
&k_{t+1} = k_t + \lambda_{\text{k}}(\gamma L_{\text{real}} - L_{\text{fake}})
\end{aligned} \tag{7}$$

Where L_{D} is the discriminator loss for optimizing each pixel, k_t is introduced to keep the balance between generator and discriminator in order to prevent gradient from collapsing, λ_{k} and γ are hyper parameters. λ_{k} is equivalent to the learning rate of k and γ decides the balance ratio during the optimization. In this study, k_0 is initialized as 0, λ_{k} is set to 0.001, and γ is set to 0.5. When the discriminator overwhelms the generator, γL_{real} will be less than L_{fake} and k_t will decrease to make the term L_{real} more dominant. Thus the generator will catch up and L_{fake} will decrease to achieve the balance. It works in the same way for the opposite case.

ALGORITHM STEPS

The specific algorithm (workflow) proceeds as follows:

Step 1: Each wheat image derived from ACID database is divided into several regions, and the target ear is centered on the images with roughly the same scale;

Step 2: All image patches are wrapped to the size of 256×256 and divided into training and test samples;

Step 3: Training samples are augmented by random rotation, scaling, and flipping;

Step 4: The augmented training samples are used as input to train the proposed network model (Figure 2);

Step 5: The test samples are input into the trained networks, and the generator outputs the number and the positions of wheat ears and spikelets.

RESULTS AND DISCUSSIONS

NUMERICAL RESULTS

This study uses F1 score and recall as quantitative evaluation indexes (Pound et al., 2017). The ear F1 scores of different methods (Hang, 2016; Pound, 2017; He, 2016; Xie, 2017) during training and testing at normalized distance of 0.2 are compared and shown in Figures 4 and 5. The recall values of different methods during testing at normalized distance of 0.2 are compared and displayed in Figures 6. Different normalization distances are also compared and presented in Figure 7. In testing, VGG, ResNet50, ResNeXt50, hourglass network and the proposed method achieve 85%, 86%, 82%, 89%,and 90.5% F1 scores, and 82.3%,83.1%,82.9%,85.3%,90.8% recall values. The hourglass network, which is adopted as baseline in the proposed method, is superior to the other deep learning networks. F1 score and recall values of the proposed method improve by 1.5% and 5.5% compared with those of the hourglass network. The proposed method solves the problem of wheat ear occlusion effectively, but the occluded spikelets are not annotated in the ACID data set. Thus, the recognition accuracy of spikelets is not greatly improved.

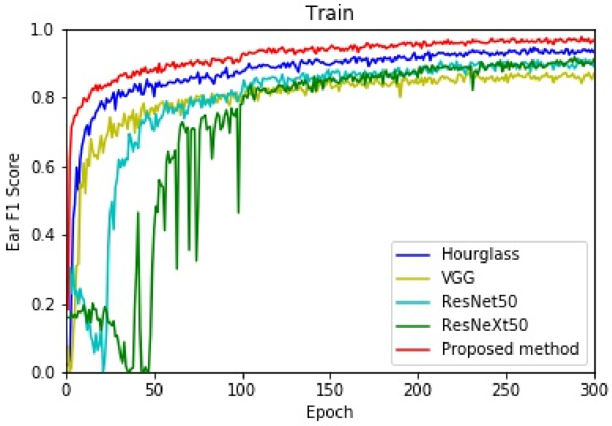

Figure 4. Comparison of ear F1 scores during training

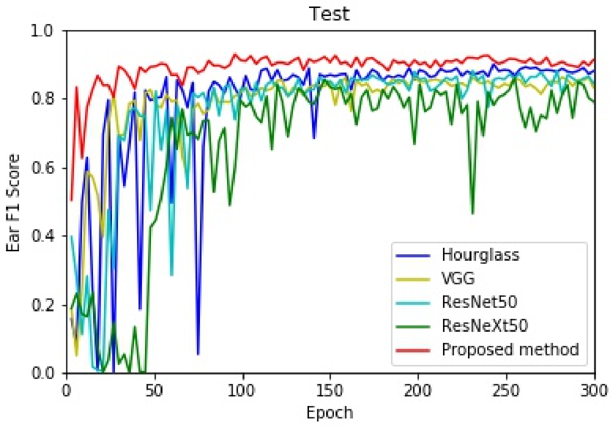

Figure 5. Comparison of ear F1 scores during testing

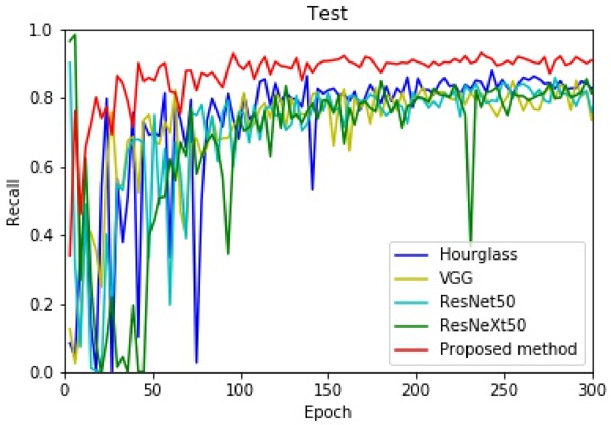

Figure 6. Comparison of ear recall during testing

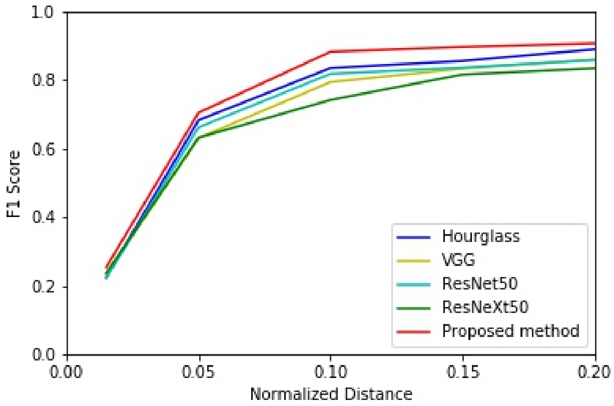

Figure 7. Comparisons of normalized distances in the ACID data set

COMPONENT ANALYSIS

The experiments use four stacked hourglass networks as baseline to evaluate the efficacy of the combination of attention mechanism and adversarial networks (Pound et al., 2017). The proposed components, namely, the attention model and the adversarial networks, are analyzed by comparing their F1 scores based on the baseline networks. Figure 8 shows the performance of the proposed components.

Attention model The holistic attention model is added to the end of each hourglass stack. The model yields 84.9% F1 score, which is 0.5% higher than the baseline networks.

Adversarial networks The self-adversarial networks are used to replace the baseline networks to evaluate their efficacy. The addition of self-adversarial networks improved the F1 score by 1.3%.

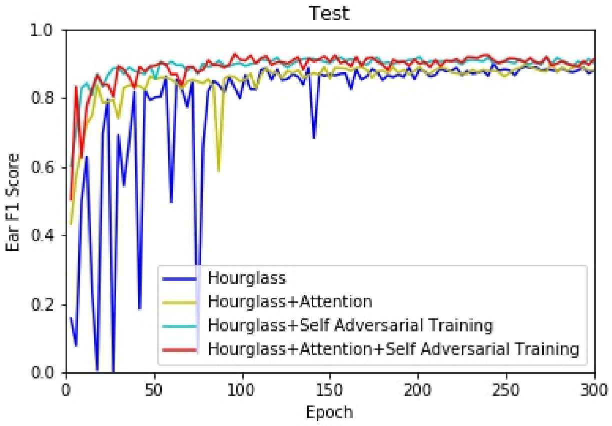

Figure 8. Comparisons of self-adversarial training, attention model and baseline network

Conclusion

In this study, attention model and adversarial networks are combined for multi-task wheat phenotyping. The adversarial networks displayed the same architecture as the generator. Serving as a critic, the adversarial networks distinguish the unreasonable locations of ears and spikelets and provide useful hints for generator to improve heatmaps. Attention model drives the proposed networks to concentrate on complex wheat ear configurations. The proposed method is evaluated using the ACID data set. The experimental results verify the effectiveness of the proposed method. The network structure that combines the attention model with the adversarial networks is also helpful in realizing other visual tasks.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 61672032. At the same time, thanks to Michael Pound for providing his data set to facilitate the next work of the paper.

References

Ba, J., Mnih, V., & Kavukcuoglu, K. (2015). Multiple object recognition with visual attention. *Int. Conf. of Learning Representation.*

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Int. Conf. of Learning Representation.*

Chou, C. J., Chien, J. T., & Chen, H. T. (2018). Self adversarial training for human pose estimation. *Asia-Pacific Signal and Information Process. Association Annual Summit and Conf.*. IEEE.

Choudhury, S. D., Stoer ger, V., Samal, A., Schnable, J. C., Liang, Z., & Yu, J. G. (2016). Automated vegetative stage phenotyping analysis of maize plants using visible light images. *KDD Workshop on Data Science for Food, Energy and Water.*

Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A. L., & Wang, X. (2017). Multi-context attention for human pose estimation. *Proc. Conf. on Computer Vision and Pattern Recognition.* IEEE.

Fernandez-Gallego, J. A., Kefauver, S. C., Gutiérrez, N. A., Nieto-Taladriz, M. T., & Araus, J. L. (2018). Wheat ear counting in-field conditions: high throughput and low-cost approach using RGB images. *Plant Methods*, 14(1), 22.

Fuentes, A., Yoon, S., Kim, S., & Park, D. (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*, 17(9), 2022.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. *Int. Conf. on Neural Information Process. Systems*. MIT Press.

Hang, S. T., Tatsuma, A., & Aono, M. (2016). Bluefield (KDE TUT) at LifeCLEF 2016 Plant Identification Task. *Conf. and Labs of the Evaluation Forum.*

Hu, J., Chen, Z., Yang, M., Zhang, R., & Cui, Y. (2018). A multiscale fusion convolutional neural network for plant leaf recognition. *IEEE Signal Process. Letters*, 25(6), 853-857.

Hu, G., Wu, H., Zhang, Y., Wan, M. (2019). A low shot learning method for tea leaf's disease identification. *Comput. Electron. Agric.*, 163, 104852.

Kuen, J., Wang, Z., & Wang, G. (2016). Recurrent attentional networks for saliency detection. *Proc. Conf. on Computer Vision and Pattern Recognition*. IEEE.

Luc, P., Couprie, C., Chintala, S., & Verbeek, J. (2016). Semantic segmentation using adversarial networks. *Conf. and Workshop on Neural Information Process. Systems*.

Perrin, G., Descombes, X., & Zerubia, J. (2005, September). A marked point process model for tree crown extraction in plantations. *Int. Conf. on Image Process.* IEEE.

Pound, M. P., Atkinson, J. A., Wells, D. M., Pridmore, T. P., & French, A. P.(2017). Deep learning for multi-task plant phenotyping. *Proc. Int. Conf. on Computer Vision*. IEEE.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *Computer Science*.

Shi, Y., Wang, N., Taylor, R. K., Raun, W. R., & Hardin, J. A. (2013). Automatic corn plant location and spacing measurement using laser line-scan technique. *Precision Agriculture*, 14(5), 478-494.

Singh, A. K., Ganapathysubramanian, B., Sarkar, S., & Singh, A. (2018). Deep learning for plant stress phenotyping: trends and future perspectives. *Trends in Plant Science*, 23(10),883-898.

Valerio Giuffrida, M., Scharr, H., & Tsaftaris, S. A. (2017). ARIGAN: synthetic Arabidopsis plants using generative adversarial network. *Proc. Int. Conf. on Computer Vision*. IEEE.

Zhou, C., Liang, D., Yang, X., Yang, H., Yue, J., & Yang, G. (2018). Wheat ears counting in field conditions based on multi-feature optimization and TWSVM. *Frontiers in Plant Science*, 9, 1024.

Zhou, H., Miao, H., Li, J., Jian, F., & Jayas, D. S. (2019). A low-resolution image restoration classifier network to identify stored-grain insects from images of sticky boards. *Comput. Electron. Agric.*, 162, 593-601.

Template for ASABE Journal Authors

Author 1

First name or initial	Middle name or initial	Surname	Suffix (Jr., III, etc.)	Role (job title, etc.)	Email (and phone for contact author)	Contact author? yes or no
Gensheng		Hu		Professor	hugs2906@sina.com	yes

Affiliation for Author 1

Organization	Address	Country	URL or other info.
National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University	No. 111 Jiulong Road, Economic and Technological Development Zone, Hefei, Anhui, China	China	http://ae.ahu.edu.cn/

Author 2 (repeat Author and Affiliation tables for each author)

First name or initial	Middle name or initial	Surname	Suffix (Jr., III, etc.)	Role (job title, etc.)	Email (and phone for contact author)	Contact author? yes or no
Lidong		Qian		Graduate Student	2478008518@qq.com	no

Affiliation for Author 2

Organization	Address	Country	URL or other info.
School of Electronics and Information Engineering, Anhui University	No. 111 Jiulong Road, Economic and Technological Development Zone, Hefei, Anhui, China	China	http://dy.ahu.edu.cn/

Author 3 (repeat Author and Affiliation tables for each author)

First name or initial	Middle name or initial	Surname	Suffix (Jr., III, etc.)	Role (job title, etc.)	Email (and phone for contact author)	Contact author? yes or no
Dong		Liang		professor	dliang@ahu.edu.cn	no

Affiliation for Author 3

Organization	Address	Country	URL or other info.
School of Electronics and Information Engineering,	No. 111 Jiulong Road, Economic and Technological Development Zone,	China	http://dy.ahu.edu.cn/

Anhui University	Hefei, Anhui, China		

Author 4 (repeat Author and Affiliation tables for each author)

First name or initial	Middle name or initial	Surname	Suffix (Jr., III, etc.)	Role (job title, etc.)	Email (and phone for contact author)	Contact author? yes or no
Mingzhu		Wan		undergraduate	16307130110@fudan.edu.cn	no

Affiliation for Author 4

Organization	Address	Country	URL or other info.
School of Information Science and Technology, Fudan University	No. 220 Handan Road, School of Information Science and Technology, Fudan University, Shanghai, China	China	http://www.it.fudan.edu.cn/