The 1st Agriculture-Vision Challenge: Methods and Results

Mang Tik Chiu1*, Xingqian Xu1*, Kai Wang3, Jennifer Hobbs2, Naira Hovakimyan2,1, Thomas S. Huang1, Honghui Shi3,1 Yunchoa Wei1, Zilong Huang1, Alexander Schwing1, Robert Brunner1, Ivan Dozier2, Wyatt Dozier2, Karen Ghandilyan2, David Wilson2, Hyunseong Park4, Junhee Kim4,5, Sungho Kim4, Qinghui Liu6, Michael C. Kampffmeyer7, Robert Jenssen7, Arnt B. Salberg6, Alexandre Barbosa1, Rodrigo Trevisan1, Bingchen Zhao8, Shaozuo Yu8, Siwei Yang8, Yin Wang8, Hao Sheng9, Xiao Chen9, Jingyi Su10, Ram Rajagopal9, Andrew Ng9, Van Thong Huynh11, Soo-Hyung Kim11, In-Seop Na12, Ujjwal Baid13, Shubham Innani13, Prasad Dutande13, Bhakti Baheti13, Sanjay Talbar13, Jianyu Tang14

1UIUC, 2Intelinair, 3University of Oregon, 4Agency for Defense Development, South Korea, 5DGIST, South Korea, 6Norwegian Computing Center, 7UiT The Arctic University of Norway, 8Tongji University, China, 9Stanford University, 10Chegg, Inc., 11Chonnam National University, South Korea, 12Chosun University, South Korea, 13SGGS Institute of Engineering and Technology, India, 14Tsinghua University, China

Abstract

The first Agriculture-Vision Challenge aims to encourage research in developing novel and effective algorithms for agricultural pattern recognition from aerial images, especially for the semantic segmentation task associated with our challenge dataset. Around 57 participating teams from various countries compete to achieve state-of-the-art in aerial agriculture semantic segmentation. The Agriculture-Vision Challenge Dataset was employed, which comprises of 21,061 aerial and multi-spectral farmland images. This paper provides a summary of notable methods and results in the challenge. Our submission server and leaderboard will continue to open for researchers that are interested in this challenge dataset and task; the link can be found here.

1. Introduction

Vision in agriculture has begun gaining increasing attention as recent advancements in deep learning solutions for various tasks were proven successful. Areas such as medicine and aerospace [18, 1, 32, 33, 34] have benefited from the effectiveness of vision applications in their respective domains. As a result, there have been numerous efforts that aim to apply pattern recognition techniques in agriculture to increase potential yield as well as prevent losses. Nevertheless, progress in these directions have been slow [15], which can be partially attributed to the lack of datasets that encourage relevant studies.

Semantic segmentation from aerial agricultural images, as one of the major topics in agriculture-vision applications, differs from common object or aerial image segmentation tasks in several aspects. First, farmland images are usually multi-spectral, since image channels such as near-infrared and thermal inputs are extremely helpful for field anomaly detection. Second, different from common objects with clear boundaries, farmland patterns are regions with extremely irregular shapes and scales. These distinctions make aerial agricultural image semantic segmentation a uniquely challenging task with great academic and economic potentials.

Nevertheless, inspirations for agricultural semantic segmentation can be drawn from methods aimed for common object segmentation. Recent works on segmentation in general have demonstrated impressive results [31, 7, 11, 24, 14, 12, 29, 30]. For example, SPGNet [7] leverages multi-scale context modules to improve semantic segmentation performances. The DeepLab series [3, 4, 5, 6] uses atrous convolution to further expand the receptive field, which enhanced the network’s ability to capture objects at larger scales. CCNet [11] proposed a criss-cross convolution to more efficiently capture non-local features. These techniques can potentially be transferred to semantic segmentation in agricultural images to yield similar performance gains.

Motivated by the above, the first Agriculture-Vision Challenge was held to encourage research in this area. A subset of the original Agriculture-Vision dataset [8] (i.e. the Agriculture-Vision Challenge dataset) was used. The challenge dataset contains 21,061 aerial and multi-spectral farmland images captured throughout 2019 across the US. In the following sections we describe and discuss in detail the challenge, notable methods and results.
2. The Agriculture-Vision Challenge

2.1. Challenge Dataset

The first Agriculture-Vision Challenge focuses on semantic segmentation from aerial agricultural images. Six important anomaly patterns from the Agriculture-Vision dataset [8] are to be recognized, which are cloud shadow, double plant, planter skip, standing water, waterway and weed cluster. Each image is $512 \times 512$ pixel with four input channels, namely Red, Green, Blue and Near-infrared (NIR). In addition to the input channels, a boundary map and a mask are provided to indicate areas within the farmland and the valid pixels in the image respectively. In total, the challenge dataset contains 12901/4431/3729 train/val/test images respectively. Visualization of each pattern is shown in Figure 1. Note that labels in this dataset are not mutually exclusive, which means that a pixel can contain more than one pattern. As a result, a custom metric is designed to evaluate submissions.

2.1. Evaluation Metric

To accommodate for overlapping labels, we modify the conventional mean Intersection-over-Union (mIoU) metric by categorizing predictions of any label in a pixel as a correct prediction. This enables easy adaptation of typical semantic segmentation models into our agriculture challenge.

Specifically, to compute the modified mIoU, a confusion matrix $M^{c \times c}$ ($c = 7$ is the number of classes plus background) is first computed with the following rules:

For each prediction $x$ and label set $Y$ at a pixel:

1. If $x \subseteq Y$, then $M_{y,y} = M_{y,y} + 1 \ \forall y \in Y$

2. Otherwise, $M_{x,y} = M_{x,y} + 1 \ \forall y \in Y$

Finally, the modified mIoU is computed by:

$$\frac{1}{c} \sum_{c} \frac{\text{True positive}_c}{\text{Prediction}_c + \text{Target}_c - \text{True positive}_c}$$

Figure 1: RGB images of each pattern in the challenge dataset. Note that as the original Agriculture-Vision dataset [8] is updated, more patterns are gradually being included. Images best viewed with color and zoomed in.
The modified mIoU increases the reward for a correct prediction by allowing any correct predictions to count as true positives for all ground truth labels. However, it also heavily penalizes the model if the prediction does not match any of the ground truth labels.

### 2.3. Challenge Description

The first Agriculture-Vision challenge was hosted between January 27, 2020 and April 20, 2020. Around 57 teams participated in the challenge, with about 33 published result submissions. Submissions were evaluated on the challenge test set with 3729 images and ranked based on the modified mIoU.

### 3. Results and Methods

Table 1 shows the results of the first Agriculture-Vision challenge. In this section, we review some notable submission, such as their motivations and methodologies.

#### 3.1. Team DSSC
**Hyunseong Park, Junhee Kim, Sungho Kim**  
Agency for Defense Development, DGIST

Residual DenseNet [35] with Squeeze-and-Excitation blocks [10] (RD-SE) is adopted as the base model for semantic segmentation. RD-SE is based on U-Net [25] architecture that has encoder/decoder architecture as shown in Figure 2. In RD-SE, to compensate for the spatial loss which arise during the feature extraction, residual dense blocks [35] and skip connections are utilized. Also, Squeeze-and-Excitation blocks (SE block) [10] are used to recalibrate channel-wise feature responses. Five convolution layers with kernel size 3x3 and batch normalization [13] are included in one residual dense block.

![Figure 2: Team DSSC: Residual DenseNet with Expert Network architecture.](image)

Expert networks were also used to segment less frequent class objects. In this challenge, two expert networks are trained for minor classes (i.e. planter skip, double plant and background). Although expert networks are based on RD-SE, they have a lighter architecture and can be trained faster than RD-SE. Trained expert networks support RD-SE to segment minority patterns. The overall process is shown in Figure 2. From the input images, RD-SE networks produce the prediction maps. If there are pixels classified as planter skip, the expert networks are implemented to segment on the same images. The prediction results for both RD-SE and expert networks are combined to make final prediction. Unlike expert networks for planter skip, the expert networks for standing water is used when there are pixels classified as planter skip and standing water from RD-SE. The result requires several steps of post-processing, including transition from planter skip to standing water (when both labels appear in the same field), removal of small labels and morphological closing.

#### 3.2. Team SCG_Vision
**Qinghui Liu, Michael C. Kampffmeyer, Robert Jenssen, Arnt B. Salberg**  
Norwegian Computing Center, UiT The Arctic University of Norway

The proposed model uses the self-constructing graph (SCG) [20] module combined with graph convolutional network [17] for aerial agricultural semantic segmentation. Since aerial images are rotational invariant, three SCG-GCN modules are used to extract features at multiple views. The proposed model architecture is shown in Figure 3.

![Figure 3: Team SCG_Vision: Multi-view Self-Constructing Graph Convolutional Network architecture.](image)

To overcome the class imbalance problem in the challenge dataset, an adaptive class reweighing loss is designed. A positive-negative class balanced function is further adopted to accommodate for negative samples. Details of this work can be found in our workshop proceedings: *Multi-view Self-Constructing Graph Convolutional Networks with Adaptive Class Weighting Loss for Semantic Segmentation*.

#### 3.3. Team AGR
**Alexandre Barbosa, Rodrigo Trevisan**  
University of Illinois at Urbana Champaign

To avoid “overlooking” the less frequent classes during training, the concept of focal loss [19] was used for imbalanced datasets. The key idea is to dynamically scale the...
<table>
<thead>
<tr>
<th>Submission</th>
<th>modified mIoU</th>
<th>Background</th>
<th>Cloud shadow</th>
<th>Double plant</th>
<th>Planter skip</th>
<th>Standing water</th>
<th>Waterway</th>
<th>Weed cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSSC</td>
<td>63.9</td>
<td>80.6</td>
<td>56</td>
<td>57.9</td>
<td>57.5</td>
<td>75</td>
<td>63.7</td>
<td>56.9</td>
</tr>
<tr>
<td>seungjae</td>
<td>62.2</td>
<td>79.3</td>
<td>44.4</td>
<td>60.4</td>
<td>65.9</td>
<td>76.9</td>
<td>55.4</td>
<td>53.2</td>
</tr>
<tr>
<td>yjl9122</td>
<td>61.5</td>
<td>80.1</td>
<td>53.7</td>
<td>46.1</td>
<td>48.6</td>
<td>76.8</td>
<td>71.5</td>
<td>53.6</td>
</tr>
<tr>
<td>SCG_Vision</td>
<td>60.8</td>
<td>80.5</td>
<td>51</td>
<td>58.6</td>
<td>49.8</td>
<td>72</td>
<td>59.8</td>
<td>53.8</td>
</tr>
<tr>
<td>AGR</td>
<td>60.5</td>
<td>80.2</td>
<td>43.8</td>
<td>57.5</td>
<td>51.6</td>
<td>75.3</td>
<td>66.2</td>
<td>49.2</td>
</tr>
<tr>
<td>SYDu</td>
<td>59.5</td>
<td>81.3</td>
<td>41.6</td>
<td>50.3</td>
<td>43.4</td>
<td>73.2</td>
<td>71.7</td>
<td>55.2</td>
</tr>
<tr>
<td>agri</td>
<td>59.2</td>
<td>78.2</td>
<td>55.8</td>
<td>42.9</td>
<td>42</td>
<td>72</td>
<td>77.5</td>
<td>64.7</td>
</tr>
<tr>
<td>TJU</td>
<td>57.4</td>
<td>79.9</td>
<td>36.6</td>
<td>54.8</td>
<td>41.4</td>
<td>69.8</td>
<td>66.9</td>
<td>52</td>
</tr>
<tr>
<td>celery030</td>
<td>55.4</td>
<td>79.1</td>
<td>38.9</td>
<td>43.3</td>
<td>41.2</td>
<td>73</td>
<td>61.5</td>
<td>50.5</td>
</tr>
<tr>
<td>stevenwudi</td>
<td>55</td>
<td>77.4</td>
<td>42</td>
<td>54.4</td>
<td>20.1</td>
<td>69.5</td>
<td>67.7</td>
<td>53.8</td>
</tr>
<tr>
<td>PAII</td>
<td>55</td>
<td>79.9</td>
<td>38.6</td>
<td>47.6</td>
<td>26.2</td>
<td>74.6</td>
<td>62.1</td>
<td>55.7</td>
</tr>
<tr>
<td>agrichallenge12</td>
<td>54.6</td>
<td>80.9</td>
<td>50.9</td>
<td>39.3</td>
<td>29.2</td>
<td>73.4</td>
<td>57.8</td>
<td>50.5</td>
</tr>
<tr>
<td>hui</td>
<td>54</td>
<td>80.2</td>
<td>41.6</td>
<td>46.4</td>
<td>20.8</td>
<td>72.8</td>
<td>64.8</td>
<td>51.4</td>
</tr>
<tr>
<td>shenchen616</td>
<td>53.7</td>
<td>79.4</td>
<td>36.7</td>
<td>56.3</td>
<td>21.6</td>
<td>67</td>
<td>61.8</td>
<td>52.8</td>
</tr>
<tr>
<td>NTU</td>
<td>53.6</td>
<td>79.8</td>
<td>41.4</td>
<td>49.4</td>
<td>13.5</td>
<td>73.3</td>
<td>61.8</td>
<td>56</td>
</tr>
<tr>
<td>tpyss</td>
<td>53</td>
<td>81.1</td>
<td>50.5</td>
<td>37.1</td>
<td>25.9</td>
<td>67.4</td>
<td>58.7</td>
<td>50.1</td>
</tr>
<tr>
<td>Simple</td>
<td>52.7</td>
<td>80.2</td>
<td>40</td>
<td>45.2</td>
<td>24.6</td>
<td>70.9</td>
<td>57.6</td>
<td>50.4</td>
</tr>
<tr>
<td>Ursus</td>
<td>52.3</td>
<td>78.9</td>
<td>36.3</td>
<td>37.8</td>
<td>34.4</td>
<td>69.3</td>
<td>57.1</td>
<td>52.3</td>
</tr>
<tr>
<td>liepieshov</td>
<td>52.1</td>
<td>77.2</td>
<td>40.2</td>
<td>46</td>
<td>16</td>
<td>71.3</td>
<td>62.9</td>
<td>51.1</td>
</tr>
<tr>
<td>Lunhao</td>
<td>49.4</td>
<td>79.5</td>
<td>40.4</td>
<td>38.8</td>
<td>10.5</td>
<td>69.4</td>
<td>58.3</td>
<td>49.1</td>
</tr>
<tr>
<td>tetelias-mipt</td>
<td>49.2</td>
<td>80.4</td>
<td>37.8</td>
<td>34.8</td>
<td>4.6</td>
<td>70.6</td>
<td>62.5</td>
<td>53.8</td>
</tr>
<tr>
<td>Dataloader</td>
<td>48.9</td>
<td>79.1</td>
<td>42</td>
<td>35.8</td>
<td>9.1</td>
<td>68.7</td>
<td>56.7</td>
<td>51.3</td>
</tr>
<tr>
<td>Hakjin</td>
<td>46.4</td>
<td>78.6</td>
<td>32</td>
<td>38.3</td>
<td>1.8</td>
<td>66.2</td>
<td>58</td>
<td>49.9</td>
</tr>
<tr>
<td>JianyuTANG</td>
<td>44.6</td>
<td>78.1</td>
<td>37.9</td>
<td>31.8</td>
<td>15.4</td>
<td>47.3</td>
<td>54.8</td>
<td>46.9</td>
</tr>
<tr>
<td>Haosssr</td>
<td>43.9</td>
<td>79.2</td>
<td>21.4</td>
<td>28.1</td>
<td>2.7</td>
<td>67.5</td>
<td>56.4</td>
<td>52.3</td>
</tr>
<tr>
<td>rpartsey</td>
<td>41.5</td>
<td>72.5</td>
<td>21.6</td>
<td>36.2</td>
<td>9.1</td>
<td>59.7</td>
<td>40.7</td>
<td>50.6</td>
</tr>
<tr>
<td>TeamTiger</td>
<td>40.8</td>
<td>75.2</td>
<td>26.1</td>
<td>40.1</td>
<td>9.9</td>
<td>48</td>
<td>37.1</td>
<td>49.5</td>
</tr>
<tr>
<td>Chaturral</td>
<td>40.7</td>
<td>77.7</td>
<td>23</td>
<td>20.4</td>
<td>5</td>
<td>55</td>
<td>51</td>
<td>52.9</td>
</tr>
<tr>
<td>Sciforce</td>
<td>40.2</td>
<td>80.5</td>
<td>29.6</td>
<td>24.4</td>
<td>0</td>
<td>41.2</td>
<td>55.9</td>
<td>50</td>
</tr>
<tr>
<td>MustafaA</td>
<td>40.1</td>
<td>76.5</td>
<td>34.4</td>
<td>25.6</td>
<td>11.1</td>
<td>46</td>
<td>36.5</td>
<td>50.3</td>
</tr>
<tr>
<td>HaotianYan</td>
<td>36.8</td>
<td>77.1</td>
<td>21.9</td>
<td>25.1</td>
<td>13.7</td>
<td>57.5</td>
<td>24.3</td>
<td>37.9</td>
</tr>
<tr>
<td>gro</td>
<td>36.3</td>
<td>76.4</td>
<td>37.5</td>
<td>8.4</td>
<td>0</td>
<td>60.3</td>
<td>29.7</td>
<td>41.8</td>
</tr>
<tr>
<td>oscmansan</td>
<td>35.5</td>
<td>71.6</td>
<td>29.6</td>
<td>3</td>
<td>0</td>
<td>52.4</td>
<td>46.2</td>
<td>45.9</td>
</tr>
<tr>
<td>ThorstenC</td>
<td>33.6</td>
<td>72.3</td>
<td>22.3</td>
<td>10</td>
<td>2</td>
<td>40.8</td>
<td>40.1</td>
<td>47.8</td>
</tr>
<tr>
<td>ZHWang</td>
<td>33.5</td>
<td>76.5</td>
<td>32.4</td>
<td>12.9</td>
<td>0</td>
<td>57.2</td>
<td>15.9</td>
<td>39.9</td>
</tr>
<tr>
<td>fayzur20</td>
<td>22.1</td>
<td>65.4</td>
<td>21.8</td>
<td>2.2</td>
<td>0.2</td>
<td>23.3</td>
<td>13.4</td>
<td>28.7</td>
</tr>
<tr>
<td>gaslen2</td>
<td>21.5</td>
<td>71</td>
<td>3.3</td>
<td>17.9</td>
<td>0.8</td>
<td>10.2</td>
<td>6.9</td>
<td>40.1</td>
</tr>
<tr>
<td>dvkhandelwal</td>
<td>16.3</td>
<td>71.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>42.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ajeetsinghittd</td>
<td>10.3</td>
<td>56.9</td>
<td>0.2</td>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>14.5</td>
</tr>
</tbody>
</table>

Table 1: Challenge results ranked by modified mIoU.

cross-entropy loss according to the confidence of the prediction of each class. In addition to the focal loss, the Lovász-Softmax [2] function was added, which is shown to be a good surrogate for the intersection-over-union metric used to evaluate the model’s performance [2]. Initial tests suggest that using equal weights to combine the focal loss with the Lovász-Softmax loss yields better results.

Two additional input channels were tested and used in the model. The first channel contains the image’s Normalized Difference Vegetation Index (NDVI). The second additional channel used in our work is the image mask. Although pixels outside the valid mask off the image are not considered in the loss function and are not evaluated, they bring relevant information since some classes are spatially correlated with the presence of a non-valid pixel (e.g. waterways are usually marked on the border of the image mask).
The base model used is the ESP Net V2 which is a computationally efficient encoder-decoder network. The model was trained from random initialization of its weights, Adam optimizer. Dropout layers were introduced with a probability of 0.5. The training converged on average in about 35 epochs. The final submission is trained over both training and validation set.

### 3.4. Team TJU
Bingchen Zhao, Shaozuo Yu, Siwei Yang, Yin Wang
Tongji University

In the proposed model, switchable normalization modules are incorporated with the IBN-Net to allow efficient data fusion and reduce feature divergence. Figure 4 shows the proposed module. The proposed method aims at resolving the divergence caused by appearance differences between RGB imagery and Near-infrared inputs present in the challenge dataset. In addition, due to potential overlaps of labels in the dataset, the problem is treated as independent binary segmentation tasks for each label type. The Lovász hinge loss is used to directly optimize on IoU. Details of this work can be found in our workshop proceedings: Reducing the feature divergence of RGB and near-infrared images using Switchable Normalization.

![Figure 4: Team TJU: the proposed IBN-s block.](image)

### 3.5. Team Haossr
Hao Sheng, Xiao Chen, Jingyi Su, Ram Rajagopal, Andrew Ng
Stanford University, Chegg, Inc.

This work focuses on exploring effective fusion techniques for multi-spectral agricultural images. A generalized vegetation index is proposed that is learnable by deep neural networks. The generalized vegetation index module learns a vegetation index feature map given multi-spectral inputs, which can be concatenated with the original color channels and fed into a deep network for inference. In addition, an additive group normalization module is introduced to smoothly train the proposed model with the generalized vegetation index output. An illustration of the fusion module is shown in Figure 5. Details of this work can be found in our workshop proceedings: Effective Data Fusion with Generalized Vegetation Index: Evidence from Land Cover Segmentation in Agriculture.

![Figure 5: Team Haossr: illustration of the fusion module for the generalized vegetation index.](image)

### 3.6. Team CNUPR.TH2L
Van Thong Huynh, Soo-Hyung Kim, In-Seop Na
Chonnam National University, Chosun University

A Deep Convolutional Encoder-Decoder architecture is deployed to segment the aerial farmland images. The encoder is based on MobileNetV2 with an attention block to assign the contribution of each spectral channel. In the decoder module, ASPP blocks are utilized and squeeze-excitation blocks are used to upsample the feature map to the original input size. An overview of the method is shown in Figure 6.

![Figure 6: Team CNUPR.TH2L: pipeline.](image)

The network is built with Keras in Tensorflow 2.1 and trained with SGD optimizer. Data augmentation is per-
formed by random flip and/or 90 degree rotation on each image except images that contain only weed clusters. This leads to 38731 images in training set. 6400 images are randomly selected in the training set to optimize the networks in each epoch. A learning rate from 0.05 to 0.3 is used with the cyclical scheduler [27]. Due to the highly imbalanced labels in the dataset, class-balanced weighting [9] is used with focal loss [19] as objective function. The source code of the method is available at https://github.com/th2l/Agriculture-Vision-Segmentation.

3.7. Team TeamTiger
Ujjwal Baid, Shubham Innani, Prasad Dutande, Bhakti Baheti, Sanjay Talbar
SGGS Institute of Engineering and Technology

The following challenges were incurred for the given segmentation task, (1) Shape and size of the area covered by each anomaly pattern are different; (2) The number of images each class is different; (3) There are overlapping labels. To cope with the challenges mentioned above, an encoder-decoder architecture using EfficientNet [28] and a feature pyramid decoder is used. The proposed encoder-decoder architecture is shown in Figure 7.

Figure 7: Team TeamTiger: proposed encoder-decoder architecture.

The proposed end-to-end semantic segmentation model is built with Tensorflow 2.0 and Keras. The network is fed with $512 \times 512 \times 4$ pixel images with a batch size of four for 100 epochs. To penalize incorrect outputs from the model while training, the Jaccard loss is used with Adam [16] as the optimizer. The learning rate is kept at 0.001 for initial epochs and then decreased five times whenever the validation does not change for three consecutive epochs.

4. Conclusion
To accommodate the rapidly changing computer vision technique in agriculture, the first Agriculture-Vision Challenge targets on efficiently and accurately recognizing several important field patterns from aerial images through semantic segmentation paradigm. Approximately 57 teams around the globe participate in this competition in which 7 leading teams, together with their novel methods, are selected for this paper. Yet our vision of agriculture should be extended beyond segmentation. The inclusive topics about agriculture have initiated many new platforms for future computer vision researches. Therefore, we can expect that, in the near future, more challenging agriculture applications will be brought out, and more powerful computer vision techniques will be developed to better assist these applications as well.

References


