This CVPR 2020 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

The 1st Agriculture-Vision Challenge: Methods and Results

Mang Tik Chiu[1*], Xingqian Xu[1*], Kai Wang[3], Jennifer Hobbs[2], Naira Hovakimyan[2,1], Thomas S. Huang[1], Honghui Shi[3,1]
Yunchao Wei[1], Zilong Huang[1], Alexander Schwing[1], Robert Brunner[1],Ivan Dozier[2], Wyatt Dozier[2], Karen Ghandilyan[2],
David Wilson[2], Hyunseong Park[4], Junhee Kim[4,5], Sungho Kim[4], Qinghui Liu[6], Michael C. Kampffmeyer[7], Robert Jenssen[7],
Arnt B. Salberg[6], Alexandre Barbosa[1], Rodrigo Trevisan[1], Bingchen Zhao[8], Shaozuo Yu[8], Siwei Yang[8], Yin Wang[8], Hao Sheng[9],
Xiao Chen[9], Jingyi Su[10], Ram Rajagopal[9], Andrew Ng[9], Van Thong Huynh[11], Soo-Hyung Kim[11], In-Seop Na[12]
Ujjwal Baid[13], Shubham Innani[13], Prasad Dutande[13], Bhakti Baheti[13], Sanjay Talbar[13], Jianyu Tang[14]

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

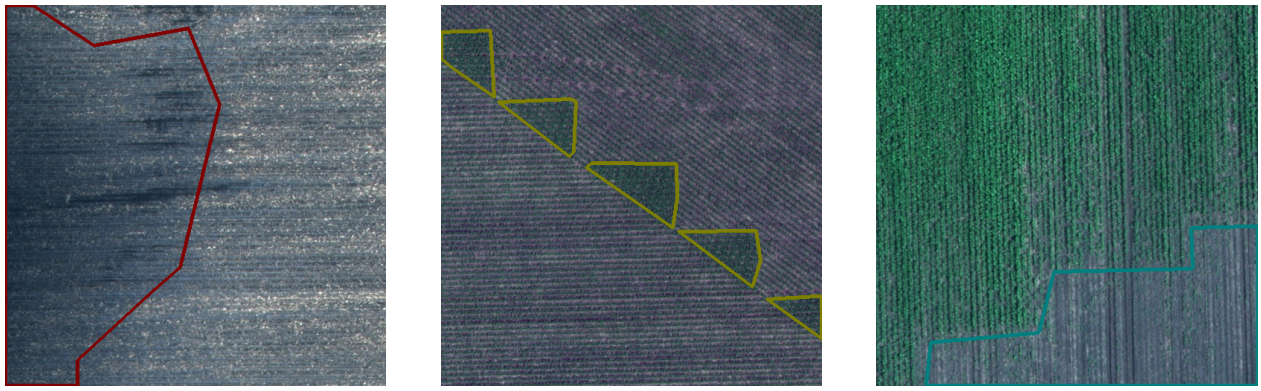
Vision in agriculture has begun gaining increasing attention as recent advancements in deep learning solutions for various tasks were proven successful. Areas such as medicine and aerospace [18, 1, 32, 33, 34] have benefited from the effectiveness of vision applications in their respective domains. As 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.

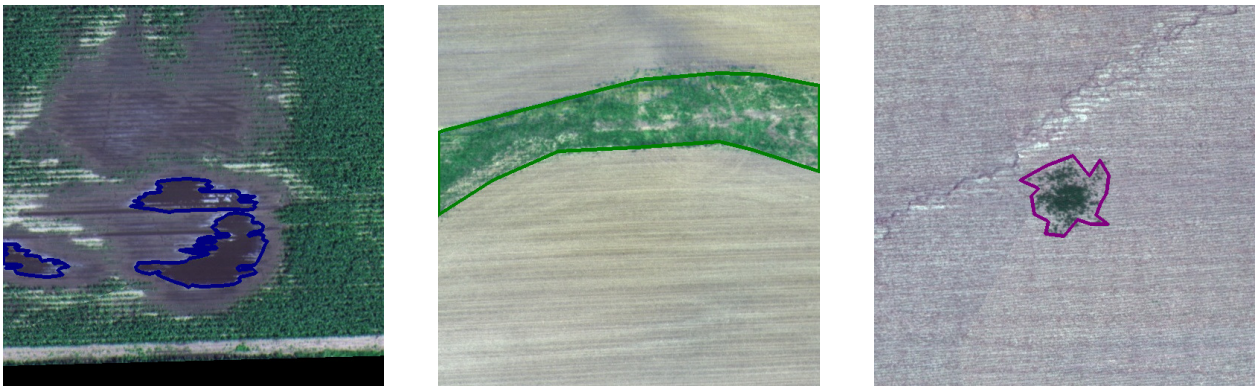
Nevertheless, inspirations for agricultural semantic segmentation can be drawn from methods aimed for common object segmentation. Recent works on segmentation in general have demonstrated impressive results [31, 7, 11, 24, 14, 12, 29, 30]. For example, SPGNet [7] leverages multi-scale context modules 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.

[*] indicates joint first author. For more information on our database and other related efforts in Agriculture-Vision, please visit our CVPR 2020 workshop and challenge website https://www.agriculture-vision.com.



(a) Cloud shadow (b) Double plant (c) Planter skip



(d) Standing water (e) Waterway (f) Weed cluster

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.

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 512×512 pixel with four input channels, namely Red, Green, Blue and Near-infrared (NIR). In addtion 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.

2.2. 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 $M_{y,y} = M_{y,y} + 1 \quad \forall y \in Y$

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

Finally, the modified mIoU is computed by:

$$\frac{1}{c} \sum_{c} \frac{True\ positive_c}{Prediction_c + Target_c - True\ positive_c}$$

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 publicized 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.

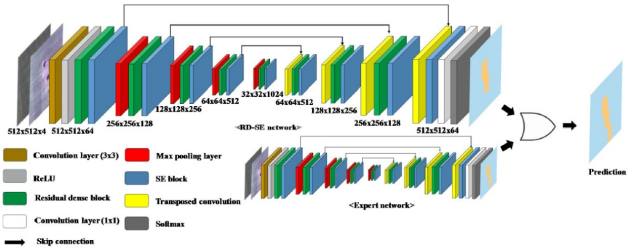


Figure 2: Team DSSC: Residual DenseNet with Expert Network architecture.

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 and standing water). The planter skip expert network takes also the double plant images as training input since many planter skip patterns also appear in the same image with double plant patterns. Therefore, the planter skip expert network considers 3 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.

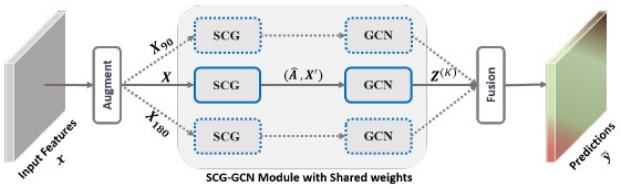


Figure 3: Team SCG_Vision: Multi-view Self-Constructing Graph Convolutional Network architecture.

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

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

Table 1: Challenge results ranked by modifeid 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 [22] network. The model was trained from random initialization of its weights, Adam optimizer [16]. 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 [21] modules are incorporated with the IBN-Net [23] 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 [2] 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**.

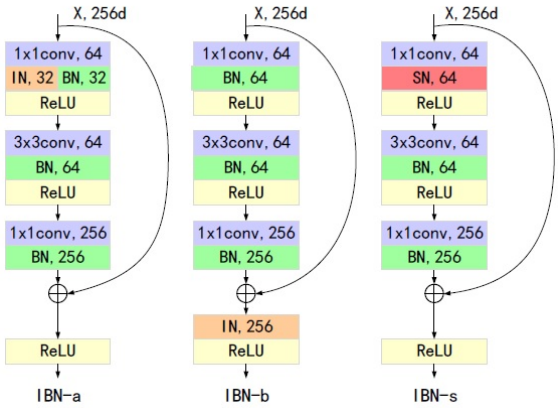


Figure 4: Team TJU: the proposed IBN-s block.

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**.

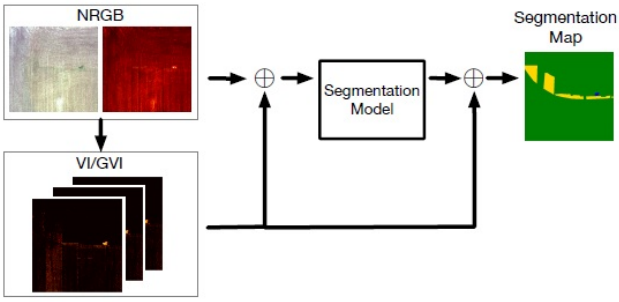


Figure 5: Team Haossr: illustration of the fusion module for the generalized vegetation index.

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 [26] with an attention block to assign the contribution of each spectral channel. In the decoder module, ASPP blocks [4] are utilized and squeeze-excitation blocks [10] are used to upsample the feature map to the original input size. An overview of the method is shown in Figure 6.

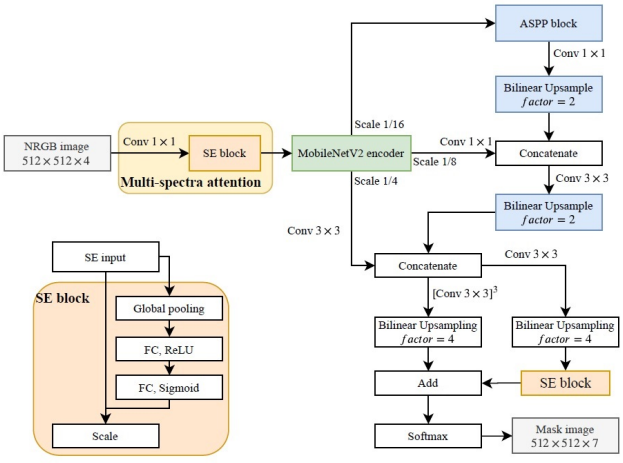


Figure 6: Team CNUPR_TH2L: pipeline.

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.

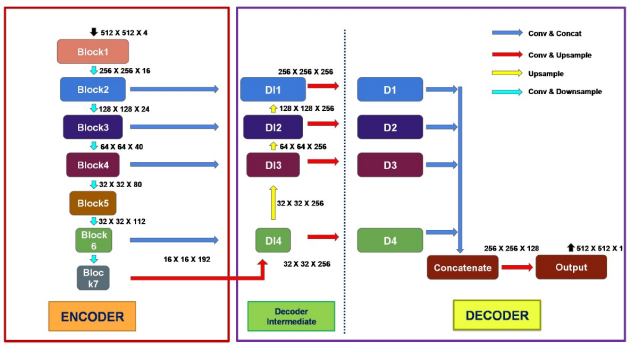


Figure 7: Team TeamTiger: proposed encoder-decoder architecture.

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

[1] AK Aniyan and Kshitij Thorat. Classifying radio galaxies with the convolutional neural network. *The Astrophysical Journal Supplement Series*, 230(2):20, 2017. 1

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018. 4, 5

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 1

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs (2016). *arXiv preprint arXiv:1606.00915*, 2016. 1, 5

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[7] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas S Huang, Wen-Mei Hwu, and Honghui Shi. Spgnet: Semantic prediction guidance for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5218–5228, 2019. 1

[8] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. *arXiv preprint arXiv:2001.01306*, 2020. 1, 2

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Com-*

puter Vision and Pattern Recognition, pages 9268–9277, 2019. 6

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3, 5

[11] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 603–612, 2019. 1

[12] Zilong Huang, Yunchao Wei, Xinggang Wang, Honghui Shi, Wenyu Liu, and Thomas S Huang. Alignseg: Feature-aligned segmentation networks. *arXiv preprint arXiv:2003.00872*, 2020. 1

[13] Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in neural information processing systems*, pages 1945–1953, 2017. 3

[14] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson WH Lau, and Thomas S Huang. Geometry-aware distillation for indoor semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2869–2878, 2019. 1

[15] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018. 1

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5, 6

[17] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3

[18] David B Larson, Matthew C Chen, Matthew P Lungren, Safwan S Halabi, Nicholas V Stence, and Curtis P Langlotz. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*, 287(1):313–322, 2017. 1

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 6

[20] Qinghui Liu, Michael Kampffmeyer, Robert Jenssen, and Arnt-Børre Salberg. Self-constructing graph convolutional networks for semantic labeling. *arXiv preprint arXiv:2003.06932*, 2020. 3

[21] Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, and Jingyu Li. Differentiable learning-to-normalize via switchable normalization. *arXiv preprint arXiv:1806.10779*, 2018. 5

[22] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018. 5

[23] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 5

[24] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8843–8850, 2019. 1

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5

[27] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017. 6

[28] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 6

[29] Zhonghao Wang, Yunchao Wei, Rogerior Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S Huang, and Honghui Shi. Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. *arXiv preprint arXiv:2004.00794*, 2020. 1

[30] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerior Feris, Jinjun Xiong, Wen mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. *arXiv preprint arXiv:2003.08040*, 2020. 1

[31] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. 1

[32] Hanchao Yu, Yang Fu, Haichao Yu, Yunchao Wei, Xinchao Wang, Jianbo Jiao, Matthew Bramlet, Thenkurussi Kesavadas, Honghui Shi, Zhangyang Wang, et al. A novel framework for 3d-2d vertebra matching. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 121–126. IEEE, 2019. 1

[33] Haichao Yu, Ding Liu, Honghui Shi, Hanchao Yu, Zhangyang Wang, Xinchao Wang, Brent Cross, Matthew Bramler, and Thomas S Huang. Computed tomography super-resolution using convolutional neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3944–3948. IEEE, 2017. 1

[34] Hanchao Yu, Shanhui Sun, Haichao Yu, Xiao Chen, Honghui Shi, Thomas Huang, and Terrence Chen. Foal: Fast online adaptive learning for cardiac motion estimation. *arXiv preprint arXiv:2003.04492*, 2020. 1

[35] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 3