Agriculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analysis

Mang Tik Chiu[1*], Xingqian Xu[1*], Yunchao Wei[1], Zilong Huang[1],
Alexander Schwing[1], Robert Brunner[1], Hrant Khachatrian[2], Hovnatan Karapetyan[2],
Ivan Dozier[2], Greg Rose[2], David Wilson[2], Adrian Tudor[2], Naira Hovakimyan[2,1],
Thomas S. Huang[1], Honghui Shi[3,1]

[1]UIUC, [2]Intelinair, [3]University of Oregon

Abstract

The success of deep learning in visual recognition tasks has driven advancements in multiple fields of research. Particularly, increasing attention has been drawn towards its application in agriculture. Nevertheless, while visual pattern recognition on farmlands carries enormous economic values, little progress has been made to merge computer vision and crop sciences due to the lack of suitable agricultural image datasets. Meanwhile, problems in agriculture also pose new challenges in computer vision. For example, semantic segmentation of aerial farmland images requires inference over extremely large-size images with extreme annotation sparsity. These challenges are not present in most of the common object datasets, and we show that they are more challenging than many other aerial image datasets. To encourage research in computer vision for agriculture, we present ***Agriculture-Vision****: a large-scale aerial farmland image dataset for semantic segmentation of agricultural patterns. We collected* $94,986$ *high-quality aerial images from* $3,432$ *farmlands across the US, where each image consists of RGB and Near-infrared (NIR) channels with resolution as high as 10 cm per pixel. We annotate nine types of field anomaly patterns that are most important to farmers. As a pilot study of aerial agricultural semantic segmentation, we perform comprehensive experiments using popular semantic segmentation models; we also propose an effective model designed for aerial agricultural pattern recognition. Our experiments demonstrate several challenges Agriculture-Vision poses to both the computer vision and agriculture communities. Future versions of this dataset will include even more aerial images, anomaly patterns and image channels.*

[*] 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.

1. Introduction

Since the introduction of ImageNet [14], a large-scale image classification dataset, research in computer vision and pattern recognition using deep neural nets has seen unprecedented development [31, 23, 49, 48, 25]. Deep neural networks based algorithms have proven to be effective across multiple domains such as medicine and astronomy [34, 2, 59], across multiple datasets [20, 51, 17], across different computer vision tasks [58, 28, 56, 9, 11, 10, 47, 43, 59] and across different numerical precision and hardware architectures [57, 61]. However, progress of visual pattern recognition in agriculture, one of the fundamental aspects of the human race, has been relatively slow [29]. This is partially due to the lack of relevant datasets that encourage the study of agricultural imagery and visual patterns, which poses many distinctive characteristics.

A major direction of visual recognition in agriculture is aerial image semantic segmentation. Solving this problem is important because it has tremendous economic potential. Specifically, efficient algorithms for detecting field conditions enable timely actions to prevent major losses or to increase potential yield throughout the growing season. However, this is much more challenging compared to typical semantic segmentation tasks on other aerial image datasets. For example, to segment weed patterns in aerial farmland images, the algorithm must be able to identify sparse weed clusters of vastly different shapes and coverages. In addition, some of these aerial images have sizes exceeding 20000×30000 pixels, these images pose a huge problem for end-to-end segmentation in terms of computation power and memory consumption. Agricultural data are also inherently multi-modal, where information such as field temperature and near-infrared signal are essential for determining field conditions. These properties deviate from those of conventional semantic segmentation tasks, thus reducing their applicability to this area of research.

Dataset	# Images	# Classes	# Labels	Tasks	Image Size (pixels)	# Pixels	Channels	Resolution (GSD)
Aerial images								
Inria Aerial Image [38]	180	2	180	seg.	5000×5000	4.5B	RGB	30 cm/px
DOTA [54]	2,806	14	188,282	det.	$\leq 4000 \times 4000$	44.9B	RGB	various
iSAID [52]	2,806	15	655,451	seg.	$\leq 4000 \times 4000$	44.9B	RGB	various
AID [55]	10,000	30	10,000	cls.	600×600	3.6B	RGB	50-800 cm/px
DeepGlobe Building [13]	24,586	2	302,701	det. / seg.	650×650	10.4B	9 bands	31-124 cm/px
EuroSAT [24]	27,000	10	27,000	cls.	256×256	1.77B	13 Bands	30 cm/px
SAT-4 [3]	500,000	4	500,000	cls.	28×28	0.39B	RGB, NIR	600 cm/px
SAT-6 [3]	405,000	6	405,000	cls.	28×28	0.32B	RGB, NIR	600 cm/px
Agricultural images								
Crop/Weed discrimination [22]	60	2	494	seg.	1296×966	0.08B	RGB	N/A
Sensefly Crop Field [1]	5,260	N/A	N/A	N/A	N/A	N/A	NRG, Red edge	12.13 cm/px
DeepWeeds [42]	17,509	1†	17,509	cls.	1920×1200	40.3B	RGB	N/A
Agriculture-Vision (ours)	**94,986**	**9**	**169,086**	seg.	512×512	22.6B	**RGB, NIR**	**10/15/20 cm/px**

† DeepWeeds has only weed annotations at image-level, but there are 8 sub-categories of weeds.

Table 1: This table shows the statistics from other datasets. All datasets are compared on number of images, categories, annotations, image size, pixel numbers and color channels. If it is an aerial image dataset, we also provide the ground sample resolution (GSD). "cls.", "det." and "seg." stand for classification, detection and segmentation respectively.

To encourage research on this challenging task, we present Agriculture-Vision, a large-scale and high-quality dataset of aerial farmland images for advancing studies of agricultural semantic segmentation. We collected images throughout the growing seasons at numerous farming locations in the US, where several important field patterns were annotated by agronomy experts.

Agriculture-Vision differs significantly from other image datasets in the following aspects: (1) unprecedented aerial image resolutions up to 10 cm per pixel (cm/px); (2) multiple aligned image channels beyond RGB; (3) challenging annotations of multiple agricultural anomaly patterns; (4) precise annotations from professional agronomists with a strict quality assurance process; and (5) large size and shape variations of annotations. These features make Agriculture-Vision a unique image dataset that poses new challenges for semantic segmentation in aerial agricultural images.

Our main contributions are summarized as follows:

- We introduce a large-scale and high quality aerial agricultural image database for advancing research in agricultural pattern analysis and semantic segmentation.
- We perform a pilot study with extensive experiments on the proposed database and provide a baseline for semantic segmentation using deep learning approaches to encourage further research.

2. Related Work

Most segmentation datasets primarily focus on common objects or street views. For example, Pascal VOC [16], MS-COCO [36] and ADE20K [64] segmentation datasets respectively consist of 20, 91 and 150 daily object categories such as airplane, person, computer, etc. The Cityscapes dataset [12], where dense annotations of street scenes are available, opened up research directions in street-view scene parsing and encouraged more research efforts in this area.

Aerial image visual recognition has also gained increasing attention. Unlike daily scenes, aerial images are often significantly larger in image sizes. For example, the DOTA dataset [54] contains images with sizes up to 4000×4000 pixels, which are significantly larger than those in common object datasets at around 500×500 pixels. Yet, aerial images are often of much lower resolutions. Precisely, the CVPR DeepGlobe2018 Building Extraction Challenge [13] uses aerial images at a resolution of 31 cm/px or lower. As a result, finer object details such as shape and texture are lost and have to be omitted in later studies.

Table 1 summarizes the statistics of the most related datasets, including those of aerial images and agricultural images. As can be seen from the table, there has been an apparent lack of large-scale aerial agricultural image databases, which, in some sense, hinders agricultural visual recognition research from rapid growth as evidenced for common images [41].

Meanwhile, many agricultural studies have proposed solutions to extract meaningful information through images. These papers cover numerous subtopics, such as spectral analysis on land and crops [63, 35, 27, 30], aerial device photogrammetry [21, 32], color indices and low-level image feature analysis [50, 44, 18, 53, 15], as well as integrated image processing systems [32, 33]. One popular approach in analyzing agricultural images is to use geo-color indices such as the Normalized-Difference-Vegetation-Index (NDVI) and Excess-Green-Index (ExG). These indices have high correlation with land information such as water [60] and plantations [39]. Besides, recent papers in computer vision have been eminently motivated by deep convolu-

tional neural networks (DCNN) [31]. DCNN is also in the spotlight in agricultural vision problems such as land cover classification [37] and weed detection [40]. In a similar work [37], Lu et. al. collected aerial images using an EOS 5D camera at 650m and 500m above ground in Penzhou and Guanghan County, Sichuan, China. They labeled cultivated land vs. background using a three-layer CNN model. In another recent work [45], Rebetez et. al. utilized an experimental farmland dataset conducted by the Swiss Confederation's Agroscope research center and proposed a DCNN-HistNN hybrid model to categorize plant species on a pixel-level. Nevertheless, since their datasets are limited in scale and their research models are outdated, both works fail to fuse state-of-the-art deep learning approaches in agricultural applications in the long run.

3. The Agriculture-Vision Dataset

Agriculture-Vision aims to be a publicly available large-scale aerial agricultural image dataset that is high-resolution, multi-band, and with multiple types of patterns annotated by agronomy experts. In its current stage, we have captured 3,432 farmland images with nine types of annotations: double plant, drydown, endrow, nutrient deficiency, planter skip, storm damage, water, waterway and weed cluster. All of these patterns have substantial impacts on field conditions and the final yield. These farmland images were captured between 2017 and 2019 across multiple growing seasons in numerous farming locations in the US. The proposed Agriculture-Vision dataset contains 94,986 images sampled from these farmlands. In this section, we describe the details on how we construct the Agriculture-Vision dataset, including image acquisition, preprocessing, pattern annotation, and finally image sample generation.

3.1. Field Image Acquisition

Farmland images in the Agriculture-Vision dataset were captured by specialized mounted cameras on aerial vehicles flown over numerous fields in the US, which primarily consist of corn and soybean fields around Illinois and Iowa. All images in the current version of Agriculture-Vision were collected from the growing seasons between 2017 and 2019. Each field image contains four color channels: Near-infrared (NIR), Red, Green and Blue.

Year	Channel	Resolution	Description	Camera
2017	N, R, G, B	15cm/px	Narrow band	2×Canon SLR
2018	N, R, G	10cm/px	Narrow band	2×Nikon D850
	B	20cm/px	Wide band	1×Nikon D800E
2019	N, R, G, B	10cm/px	Narrow band	WAMS

Table 2: Camera settings for capturing the 4-channel field images: Near-infrared (N), Red (R), Green (G) and Blue (B). The Blue channel images captured in 2018 are scaled up to align with the NRG images.

The camera settings for capturing farmland images are shown in Table 2. Farmland images in 2017 were taken with two aligned Canon SLR cameras, where one captures RGB images and the other captures only the NIR channel. For farmland images in 2018, the NIR, Red and Green (NRG) channels were taken using two Nikon D850 cameras to enable 10 cm/px resolution. Custom filters were used to capture near-infrared instead of the blue channel. Meanwhile, the separate Blue channel images were captured using one Nikon D800E at 20 cm/px resolution, which were then scaled up to align with the corresponding NRG images. Farmland images in 2019 were captured using a proprietary Wide Area Multi-Spectral System (WAMS) commonly used for remote sensing. The WAMS captures all four channels simultaneously at 10 cm/px resolution. Note that compared to other aerial image datasets in Table 1, our dataset contains images in resolutions higher than all others.

3.2. Farmland image preprocessing

Farmland images captured in 2017 were already stored in regular pixel values between 0 and 255, while those captured in 2018 and 2019 were initially stored in camera raw pixel format. Following the conventional method for normalizing agricultural images, for each of the four channels in one field image, we first compute the 5^{th} and 95^{th} percentile pixel values, then clip all pixel values in the image by a lower bound and an upper bound:

$$\begin{aligned} V_{lower} &= max(0, p_5 - 0.4 \times (p_{95} - p_5)) \\ V_{upper} &= min(255, p_{95} + 0.4 \times (p_{95} - p_5)) \end{aligned} \tag{1}$$

where V_{lower}, V_{upper} stand for lower and upper bound of pixel values respectively, p_5 and p_{95} stand for the 5^{th} and 95^{th} percentile respectively.

Note that farmland images may contain invalid areas, which were initially marked with a special pixel value. Therefore, we exclude these invalid areas when computing pixel percentiles for images in 2018 and 2019.

To intuitively visualize each field image and prepare for later experiments, we separate the four channels into a regular RGB image and an additional single-channel NIR image, and store them as two JPG images.

3.3. Annotations

All annotations in Agriculture-Vision were labeled by five annotators trained by expert agronomists through a commercial software. Annotated patterns were then reviewed by the agronomists, where unsatisfactory annotations were improved. The software provides visualizations of several image channels and vegetation indices, including RGB, NIR and NDVI, where NDVI can be derived from the Red and NIR channel by:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{2}$$

Figure 1: Visualization of an aerial farmland image before sub-sampling. This image (including invalid areas, shown in black) has a size of 10875×3303 pxiels and only contains drydown annotations at the rightmost region. Due to the large image size and sparse annotation, training semantic segmentation models on entire images is impractical and inefficient.

3.4. Image sample generation

Unprocessed farmland images have extremely large image sizes. For instance, Figure 1 shows one field image with a size of 10875×3303 pixels. In fact, the largest field image we collected is 33571×24351 pixels in size. This poses significant challenges to deep network training in terms of computation time and memory consumption. In addition, Figure 1 also shows the sparsity of some annotations. This means training a segmentation model on the entire image for these patterns would be very inefficient, and would very possibly yield suboptimal results.

On the other hand, unlike common objects, visual appearances of anomaly patterns in aerial farmland images are preserved under image sub-sampling methods such as flipping and cropping. This is because these patterns represent *regions* of the anomalies instead of individual objects. As a result, we can sample image patches from these large farmland images by cropping around annotated regions in the image. This simultaneously improves data efficiency, since the proportion of annotated pixels is increased.

Motivated by the above reasons, we construct the Agriculture-Vision dataset by cropping annotations with a window size of 512×512 pixels. For field patterns smaller than the window size, we simply crop the region centered at the annotation. For field patterns larger than the window size, we employ a non-overlapping sliding window technique to cover the entirety of the annotation. Note that we discard images covered by more than 90% of annotations, such that all images retain sufficient context information.

In many cases, multiple small annotations are located near each other. Generating one image patch for every annotation would lead to severe re-sampling of those field regions, which causes biases in the dataset. To alleviate the issue, if two image patches have an Intersection-over-Union of over 30%, we discard the one with fewer pixels annotated as field patterns. When cropping large annotations using a sliding window, we also discard any image patches with only background pixels. A visualization of our sample generation method is illustrated in Figure 2, and some images in the final Agriculture-Vision dataset are shown in Figure 3.

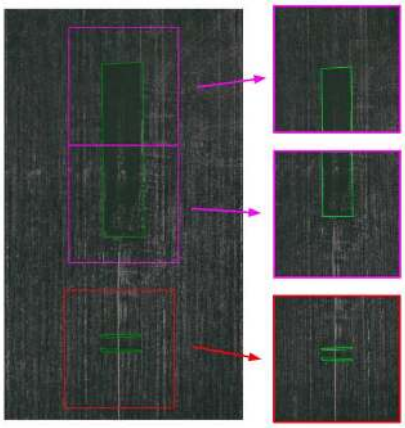
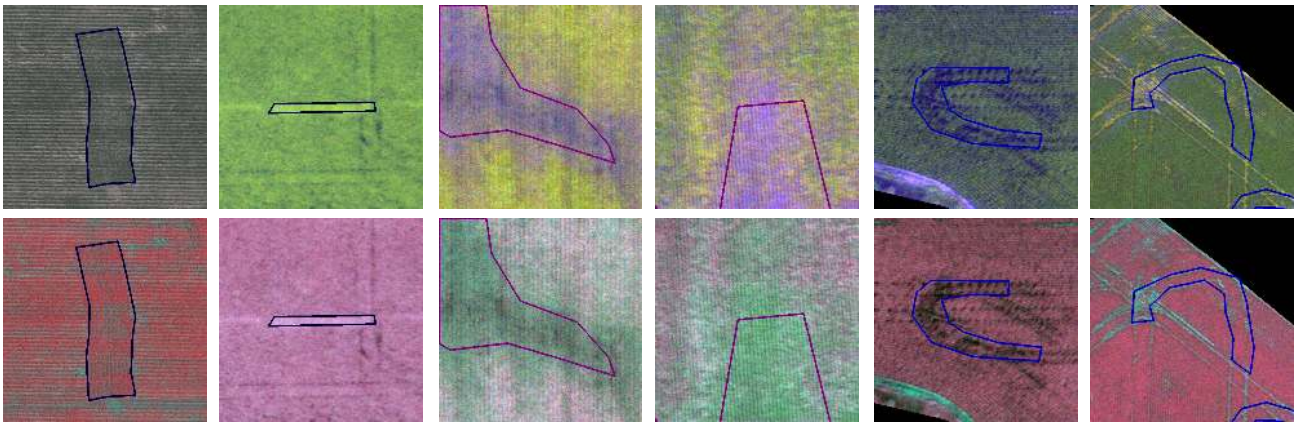


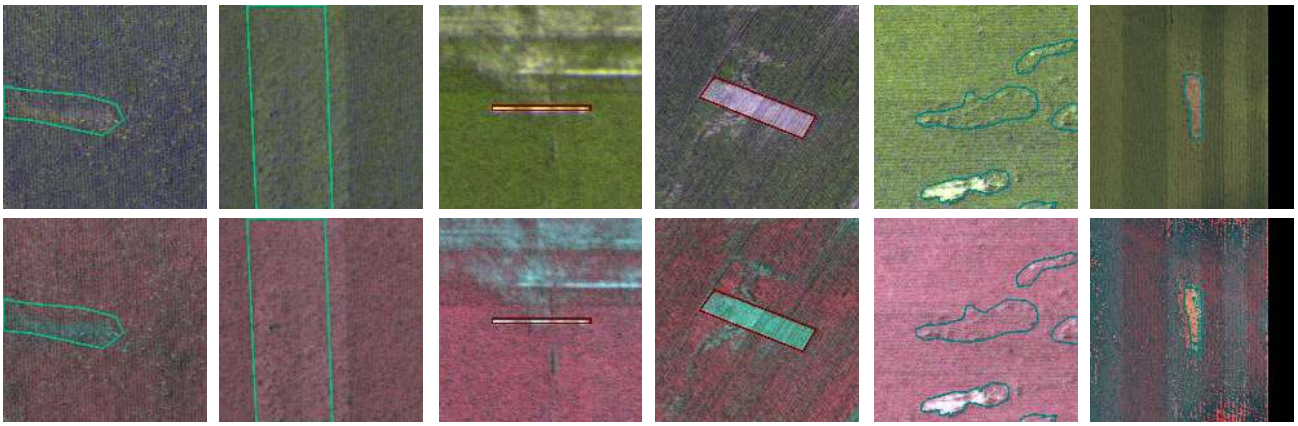
Figure 2: This figure illustrates our field image patch generation method for AgriVision. For annotations smaller than 512×512 pixels, we crop the image by a single window around the annotation center (shown in red). For larger annotations, we use multiple non-overlapping windows to cover the entire annotation (shown in purple). Note that the bottom two polygons are enclosed by just one window.

3.5. Dataset splitting

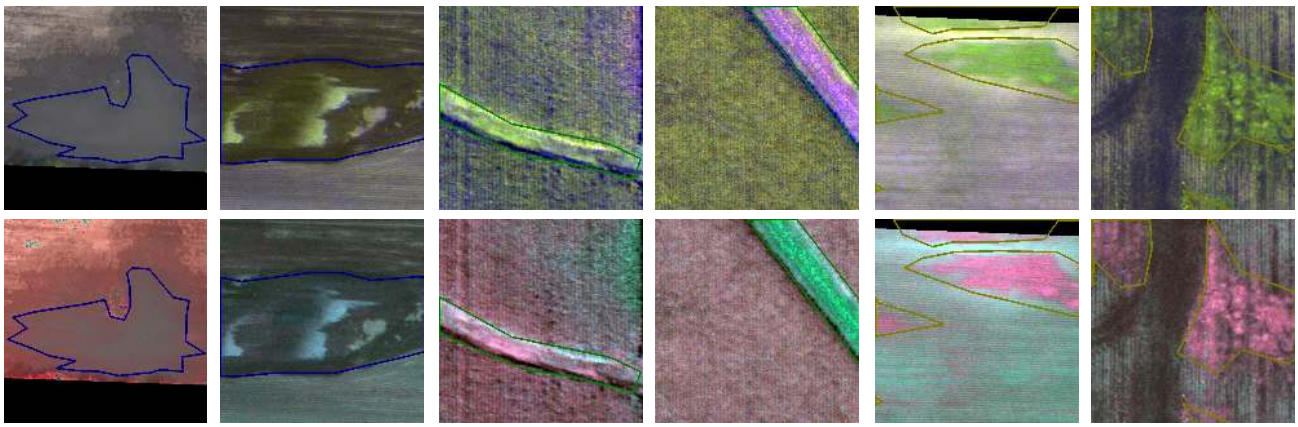
We first randomly split the 3,432 farmland images with a 6/2/2 train/val/test ratio. We then assign each sampled image to the split of the farmland image they are cropped from. This guarantees that no cropped images from the same farmland will appear in multiple splits in the final dataset. The generated Agriculture-Vision dataset thus contains 56,944/18,334/19,708 train/val/test images.



(a) Double plant (b) Drydown (c) Endrow



(d) Nutrient deficiency (e) Planter skip (f) Storm damage (not evaluated)



(g) Water (h) Waterway (i) Weed cluster

Figure 3: For each annotation, top: RGB image; bottom: NRG image. Invalid regions have been blacked out. Note the extreme size and shape variations of some annotations. Note that images in our dataset can contain mutiple patterns, the visualizations above are chosen to best illustrate each pattern. Images best viewed with color and zoomed in.

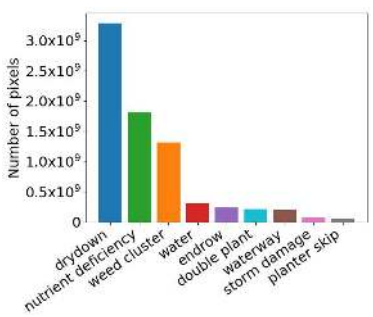


Figure 4: Total area of annotations for each class. Some categories occupy significantly larger areas than others, resulting in extreme class imbalance.

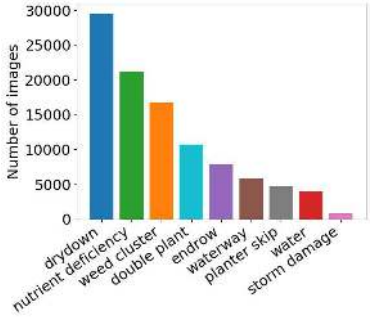


Figure 5: Number of images containing each anntoation class. A sudden drop in the number of storm damage samples indicate the difficulty for a model to recognize this pattern.

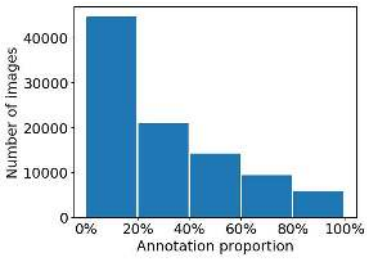


Figure 6: Percentages of annotated pixels in images. Some patterns are almost take up the entire image.

4. Dataset Statistics

4.1. Annotation areas

Field patterns have different shapes and sizes. For example, weed clusters can appear in either small patches or enormous regions, while double plant usually occur in small areas on the field. At the same time, these patterns also appear at different frequencies. Therefore, patterns that are large and more common occupy significantly larger areas than patterns that are smaller and relatively rare.

Figure 4 shows the total number of pixels for each type of annotations in Agriculture-Vision. We observe significantly more drydown, nutrient deficiency and weed cluster pixels than other categories in our dataset, which indicate extreme label imbalance across categories.

4.2. Annotation counts

The frequency at which a model observes a pattern during training determines the model's ability to recognize the same pattern during inference. It is therefore very important to understand the sample distribution for each of these field patterns in the Agriculture-Vision dataset.

Figure 5 shows the number of images that contain each annotation category. While most annotations fall under a natural and smooth occurrence distribution, we observe a sudden drop of images containing storm damage patterns. The extreme scarcity of storm damage annotations would be problematic for model training. As a result, we ignore any storm damage annotations when performing evaluations.

4.3. Annotation proportions

As previously described, field patterns can vary dramatically in size. Correspondingly in Agriculture-Vision, each generated image sample may also contain various proportions of annotations. We show in Figure 6 that many images contain more than 50% annotated pixels, some even occupy more than 80% of the image. Training a model to segment large patterns can be difficult, since recognition of field patterns relies heavily on the contextual information of the surrounding field.

5. Pilot Study on Agriculture-Vision

5.1. Baseline models

There are many popular models for semantic segmentation on common object datasets. For example, U-Net [46] is a light-weight model that leverages an encoder-decoder architecture for pixel-wise classification. PSPNet [62] uses spatial pooling at multiple resolutions to gather global information. DeepLab [4, 5, 6, 7] is a well-known series of deep learning models that use atrous convolutions for semantic segmentation. More recently, many new methods have been proposed and achieve state-of-the-art results on CityScapes benchmark. For example, SPGNet [8] proposes a Semantic Prediction Guidance (SPG) module which learns to re-weight the local features through the guidance from pixel-wise semantic prediction, and [26] proposes Criss-Cross Network (CCNet) for obtaining better contextual information in a more effective and efficient way. In our experiments, we perform comparative evaluations on the Agriculture-Vision dataset using DeepLabV3 and DeepLabV3+, which are two well-performing models across several semantic segmentation datasets. We also propose a specialized FPN-based model that outperforms these two milestones in Agriculture-Vision.

To couple with Agriculture-Vision, we make minor modifications on the existing DeepLabV3 and DeepLabV3+ architectures. Since Agriculture-Vision contains NRGB images, we duplicate the weights corresponding to the Red channel of the pretrained convolution layer. This gives a convolution layer with four input channels in the backbone.

Model	mIoU (%)	Background	Double plant	Drydown	Endrow	Nutrient deficiency	Planter skip	Water	Waterway	Weed cluster
DeepLabv3 (os=8) [6]	35.29	73.01	21.32	56.19	12.00	35.22	20.10	42.19	35.04	22.51
DeepLabv3+ (os=8) [7]	37.95	72.76	21.94	56.80	16.88	34.18	18.80	61.98	35.25	22.98
DeepLabv3 (os=16) [6]	41.66	**74.45**	25.77	57.91	19.15	**39.40**	24.25	72.35	36.42	25.24
DeepLabv3+ (os=16) [7]	42.27	74.32	25.62	**57.96**	21.65	38.42	29.22	73.19	**36.92**	23.16
Ours	**43.40**	74.31	**28.45**	57.43	**21.74**	38.86	**33.55**	**73.59**	34.37	**28.33**

Table 3: mIoUs and class IoUs of modified semantic segmentation models and our proposed FPN-based model on Agriculture-Vision **validation** set. Our model is customized for aerial agricultural images and perform better than all others.

Model	mIoU (%)	Background	Double plant	Drydown	Endrow	Nutrient deficiency	Planter skip	Water	Waterway	Weed cluster
DeepLabv3 (os=8) [6]	32.18	70.42	21.51	50.97	12.60	39.37	20.37	15.69	33.71	24.98
DeepLabv3+ (os=8) [7]	39.05	70.99	19.67	50.89	19.50	41.32	24.42	62.25	34.14	28.27
DeepLabv3 (os=16) [6]	42.22	**72.73**	25.15	**53.62**	20.99	43.95	24.57	70.42	38.63	29.91
DeepLabv3+ (os=16) [7]	42.42	72.50	25.99	53.57	24.10	**44.15**	24.39	70.33	37.91	28.81
Ours	**43.66**	72.55	**27.88**	52.32	**24.43**	43.79	**30.95**	**71.33**	**38.81**	**30.87**

Table 4: mIoUs and class IoUs of semantic segmentation models and our proposed model on Agriculture-Vision **test** set. The results are consistent with the validation set, where our model outperforms common object semantic segmentation models.

5.2. The proposed FPN-based model

In our FPN-based model, the encoder of the FPN is a ResNet [23]. We retain the first three residual blocks of the ResNet, and we change the last residual block (layer4) into a dilated residual block with rate=4. The modified block shares the same structure with the Deeplab series [4, 5, 6, 7]. We implement the lateral connections in the FPN decoder using two 3×3 and one 1×1 convolution layers. Each of the two 3×3 convolution layers is followed by a batch normalization layer (BN) and a leaky ReLU activation with a negative slope of 0.01. The last 1×1 convolution layer does not contain bias units. For the upsampling modules, instead of bilinear interpolation, we use a deconvolution layer with kernel size=3, stride=2 and padding=1, followed by a BN layer, leaky ReLU activation and another 1×1 convolution layer without bias. The output from each lateral connection and the corresponding upsampling module are added together, the output is then passed through two more 3×3 convolution layers with BN and leaky ReLU. Lastly, outputs from all pyramid levels are upsampled to the highest pyramid resolution using bilinear interpolation and are then concatenated. The result is passed to a 1×1 convolution layer with bias units to predict the final semantic map.

5.3. Training details

We use backbone models pretrained on ImageNet in all our experiments. We train each model for 25,000 iterations with a batch size of 40 on four RTX 2080Ti GPUs. We use SGD with a base learning rate of 0.01 and a weight decay of 5×10^{-4}. Within the 25,000 iterations, we first warm-up the training for 1,000 iterations [19], where the learning rate linearly grows from 0 to 0.01. We then train for 7,000 iterations with a constant learning rate of 0.01. We finally decrease the learning rate back to 0 with the "poly" rule [5] in the remaining 17,000 iterations.

Table 3 and Table 4 show the validation and test set results of DeepLabV3 and DeepLabV3+ with different output strides and our proposed FPN-based model. Our model consistently outperforms these semantic segmentation models in Agriculture-Vision. Hence, in the following experiments, we will use our FPN-based model for comparison studies.

5.4. Multi-spectral data and model complexity

One major study of our work is the effectiveness of training multi-spectral data in image recognition. Agriculture-Vision consists of NIR-Red-Green-Blue (NRGB) images, which is beyond many conventional image recognition tasks. Therefore, we investigate the differences in performance between semantic segmentation from multi-spectral images, including NRG and NRGB images, and regular RGB images.

We simultaneously investigate the impact of using models with different complexities. Specifically, we train our FPN-based model with ResNet-50 and ResNet-101 as backbone. We evaluate combinations of multi-spectral images and various backbones and report the results in Table 5.

5.5. Multi-scale data

Aerial farmland images contain annotations with vastly different sizes. As a result, models trained from images at different scales can result in significantly different performances. In order to justify our choice of using 512×512 windows to construct the Agriculture-Vision dataset, we additionally generate two versions of the dataset with different window sizes. The first version (Agriculture-Vision-1024) uses 1024×1024 windows to crop annotations. The second

Backbone	Channels	Val mIoU (%)	Test mIoU (%)
ResNet-50	RGB	39.28	38.26
ResNet-50	NRG	42.89	41.34
ResNet-50	NRGB	42.16	41.82
ResNet-101	RGB	40.48	39.63
ResNet-101	NRG	42.25	40.05
ResNet-101	NRGB	**43.40**	**43.66**

Table 5: mIoUs using our proposed model with various ResNet backbones and image channels.

		Val mIoU (%)			Test mIoU (%)		
		512	1024	MS	512	1024	MS
Train	512	**43.40**	**39.44**	**37.64**	**43.66**	**39.68**	**37.27**
	1024	36.33	34.37	36.16	35.01	35.27	35.87
	MS	34.16	32.45	35.67	31.17	30.72	35.77

Table 6: mIoUs of our model trained and tested on different Agriculture-Vision versions. 512: the proposed Agriculture-Vision dataset, 1024: Agriculture-Vision-1024, MS: Agriculture-Vision-MS. The model trained on the proposed dataset yields the best results across all versions.

version (Agriculture-Vision-MS) uses three window sizes: 1024×1024, 1536×1536 and 2048×2048.

In Agriculture-Vision-MS, images are cropped with the smallest window size that completely encloses the annotation. If an annotation exceeds 2048×2048 pixels, we again use the sliding window cropping method to generate multiple sub-samples. We use Agriculture-Vision-MS to evaluate if retaining the integrity of large annotations helps to improve performances. Note that this is different from conventional multi-scale inference used in common object image segmentation, since in Agriculture-Vision-MS the images are of different sizes.

We cross-evaluate models trained on each dataset version with all three versions. Results in Table 6 show that the model trained on the proposed Agriculture-Vision dataset with a 512×512 window size is the most stable and performs the best, thus justifying our dataset with the chosen image sampling method.

6. Discussion

We would like to highlight the use of Agriculture-Vision to tackle the following crucial tasks:

- **Agriculture images beyond RGB:** Deep convolutional neural networks (DCNN) are channel-wise expandable by nature. Yet few datasets promote in-depth research on such capability. We have demonstrated that aerial agricultural semantic segmentation is more effective using NRGB images rather than just RGB images. Future versions of Agriculture-Vision will also include thermal images, soil maps and topographic maps. Therefore, further studies in multi-spectral agriculture images are within our expectation.

- **Transfer learning:** Our segmentation task induces an uncommon type of transfer learning, where a model pretrained on RGB images of common objects is transferred to multi-spectral agricultural images. Although the gap between the source and target domain is tremendous, our experiments show that transfer learning remains an effective way of learning to recognize field patterns. Similar types of transfer learning are not regularly seen, but they are expected to become more popularized with Agriculture-Vision. The effectiveness of fine-tuning can be further explored, such as channel expansion in convolution layers and domain adaptation from common objects to agricultural patterns.

- **Learning from extreme image sizes:** The current version of Agriculture-Vision provides a pilot study of aerial agricultural pattern recognition with conventional image sizes. However, our multi-scale experiments show that there is still much to explore in effectively leveraging large-scale aerial images for improved performance. Using Agriculture-Vision as a starting point, we hope to initiate related research on visual recognition tasks that are generalizable to extremely large aerial farmland images. We envision future work in this direction to enable large-scale image analysis as a whole.

7. Conclusion

We introduce Agriculture-Vision, an aerial agricultural semantic segmentation dataset. We capture extremely large farmland images and provide multiple field pattern annotations. This dataset poses new challenges in agricultural semantic segmentation from aerial images. As a baseline, we provide a pilot study on Agriculture-Vision using well-known off-the-shelf semantic segmentation models and our specialized one.

In later versions, Agriculture-Vision will include more field images and patterns, as well as more image modalities, such as thermal images, soil maps and topographic maps. This would make Agriculture-Vision an even more standardized and inclusive aerial agricultural dataset. We hope this dataset will encourage more work on improving visual recognition methods for agriculture, particularly on large-scale, multi-channel aerial farmland semantic segmentation.

References

[1] Sensefly agriculture dataset. https://www.sensefly.com/education/datasets. Accessed:2018/11/16. 2

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

[3] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 37. ACM, 2015. 2

[4] 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. 6, 7

[5] 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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 6, 7

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6, 7

[7] 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. 6, 7

[8] 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. 6

[9] Bowen Cheng, Yunchao Wei, Honghui Shi, Shiyu Chang, Jinjun Xiong, and Thomas S Huang. Revisiting pre-training: An efficient training method for image classification. *arXiv preprint arXiv:1811.09347*, 2018. 1

[10] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Decoupled classification refinement: Hard false positive suppression for object detection. *arXiv preprint arXiv:1810.04002*, 2018. 1

[11] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. *arXiv preprint arXiv:1908.10357*, 2019. 1

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2

[13] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raska. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209. IEEE, 2018. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 1

[15] MS El-Faki, N Zhang, and DE Peterson. Weed detection using color machine vision. *Transactions of the ASAE*, 43(6):1969, 2000. 2

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2

[17] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6112–6121, 2019. 1

[18] Anatoly A Gitelson, Yoram J Kaufman, Robert Stark, and Don Rundquist. Novel algorithms for remote estimation of vegetation fraction. *Remote sensing of Environment*, 80(1):76–87, 2002. 2

[19] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 7

[20] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4805–4814, 2019. 1

[21] Norbert Haala, Michael Cramer, Florian Weimer, and Martin Trittler. Performance test on uav-based photogrammetric data collection. *Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 38(1/C22):7–12, 2011. 2

[22] Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *European Conference on Computer Vision*, pages 105–116. Springer, 2014. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 7

[24] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *arXiv preprint arXiv:1709.00029*, 2017. 2

[25] Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens Van Der Maaten, and Kilian Weinberger. Convolutional networks with dense connectivity. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

[26] 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. 6

[27] E Raymond Hunt, W Dean Hively, Stephen J Fujikawa, David S Linden, Craig ST Daughtry, and Greg W McCarty. Acquisition of nir-green-blue digital photographs from unmanned aircraft for crop monitoring. *Remote Sensing*, 2(1):290–305, 2010. 2

[28] 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

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

[30] Joshua Kelcey and Arko Lucieer. Sensor correction of a 6-band multispectral imaging sensor for uav remote sensing. *Remote Sensing*, 4(5):1462–1493, 2012. 2

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 3

[32] Andrea S Laliberte and Albert Rango. Image processing and classification procedures for analysis of sub-decimeter imagery acquired with an unmanned aircraft over arid rangelands. *GIScience & Remote Sensing*, 48(1):4–23, 2011. 2

[33] Ross D Lamm, David C Slaughter, and D Ken Giles. Precision weed control system for cotton. *Transactions of the ASAE*, 45(1):231, 2002. 2

[34] 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

[35] Valentine Lebourgeois, Agnès Bégué, Sylvain Labbé, Benjamin Mallavan, Laurent Prévot, and Bruno Roux. Can commercial digital cameras be used as multispectral sensors? a crop monitoring test. *Sensors*, 8(11):7300–7322, 2008. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[37] Heng Lu, Xiao Fu, Chao Liu, Long-guo Li, Yu-xin He, and Nai-wen Li. Cultivated land information extraction in uav imagery based on deep convolutional neural network and transfer learning. *Journal of Mountain Science*, 14(4):731–741, 2017. 3

[38] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017. 2

[39] JA Marchant, Hans Jørgen Andersen, and CM Onyango. Evaluation of an imaging sensor for detecting vegetation using different waveband combinations. *Computers and Electronics in Agriculture*, 32(2):101–117, 2001. 2

[40] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:41, 2017. 3

[41] Dmytro Mishkin, Nikolay Sergievskiy, and Jiri Matas. Systematic evaluation of convolution neural network advances on the imagenet. *Computer Vision and Image Understanding*, 161:11–19, 2017. 2

[42] Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1):2058, 2019. 2

[43] 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

[44] Anushree Ramanath, Saipreethi Muthusrinivasan, Yiqun Xie, Shashi Shekhar, and Bharathkumar Ramachandra. Ndvi versus cnn features in deep learning for land cover clasification of aerial images. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 6483–6486. IEEE, 2019. 2

[45] J Rebetez, HF Satizábal, M Mota, D Noll, L Buchi, Marina Wendling, B Cannelle, A Perez-Uribe, and Stéphane Burgos. Augmenting a convolutional neural network with local histogramsa case study in crop classification from high-resolution uav imagery. In *European Symp. on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 515–520, 2016. 3

[46] 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. 6

[47] Honghui Shi. Geometry-aware traffic flow analysis by detection and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 116–120, 2018. 1

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[49] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[50] Anju Unnikrishnan, V Sowmya, and KP Soman. Deep learning architectures for land cover classification using red and near-infrared satellite images. *Multimedia Tools and Applications*, 78(13):18379–18394, 2019. 2

[51] 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

[52] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 2

[53] David M Woebbecke, George E Meyer, K Von Bargen, and DA Mortensen. Color indices for weed identification under various soil, residue, and lighting conditions. *Transactions of the ASAE*, 38(1):259–269, 1995. 2

[54] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 2

[55] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 2

[56] 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

[57] Haichao Yu, Haoxiang Li, Honghui Shi, Thomas S. Huang, and Gang Hua. Any-precision deep neural networks. *arXiv preprint arXiv:1911.07346*, 2019. 1

[58] 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

[59] 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

[60] Pablo J Zarco-Tejada, Victoria González-Dugo, LE Williams, L Suárez, José AJ Berni, D Goldhamer, and E Fereres. A pri-based water stress index combining structural and chlorophyll effects: Assessment using diurnal narrow-band airborne imagery and the cwsi thermal index. *Remote sensing of environment*, 138:38–50, 2013. 2

[61] Xiaofan Zhang, Haoming Lu, Cong Hao, Jiachen Li, Bowen Cheng, Yuhong Li, Kyle Rupnow, Jinjun Xiong, Thomas Huang, Honghui Shi, et al. Skynet: a hardware-efficient method for object detection and tracking on embedded systems. *arXiv preprint arXiv:1909.09709*, 2019. 1

[62] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 6

[63] Xin Zhao, Yitong Yuan, Mengdie Song, Yang Ding, Fenfang Lin, Dong Liang, and Dongyan Zhang. Use of unmanned aerial vehicle imagery and deep learning unet to extract rice lodging. *Sensors*, 19(18):3859, 2019. 2

[64] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2